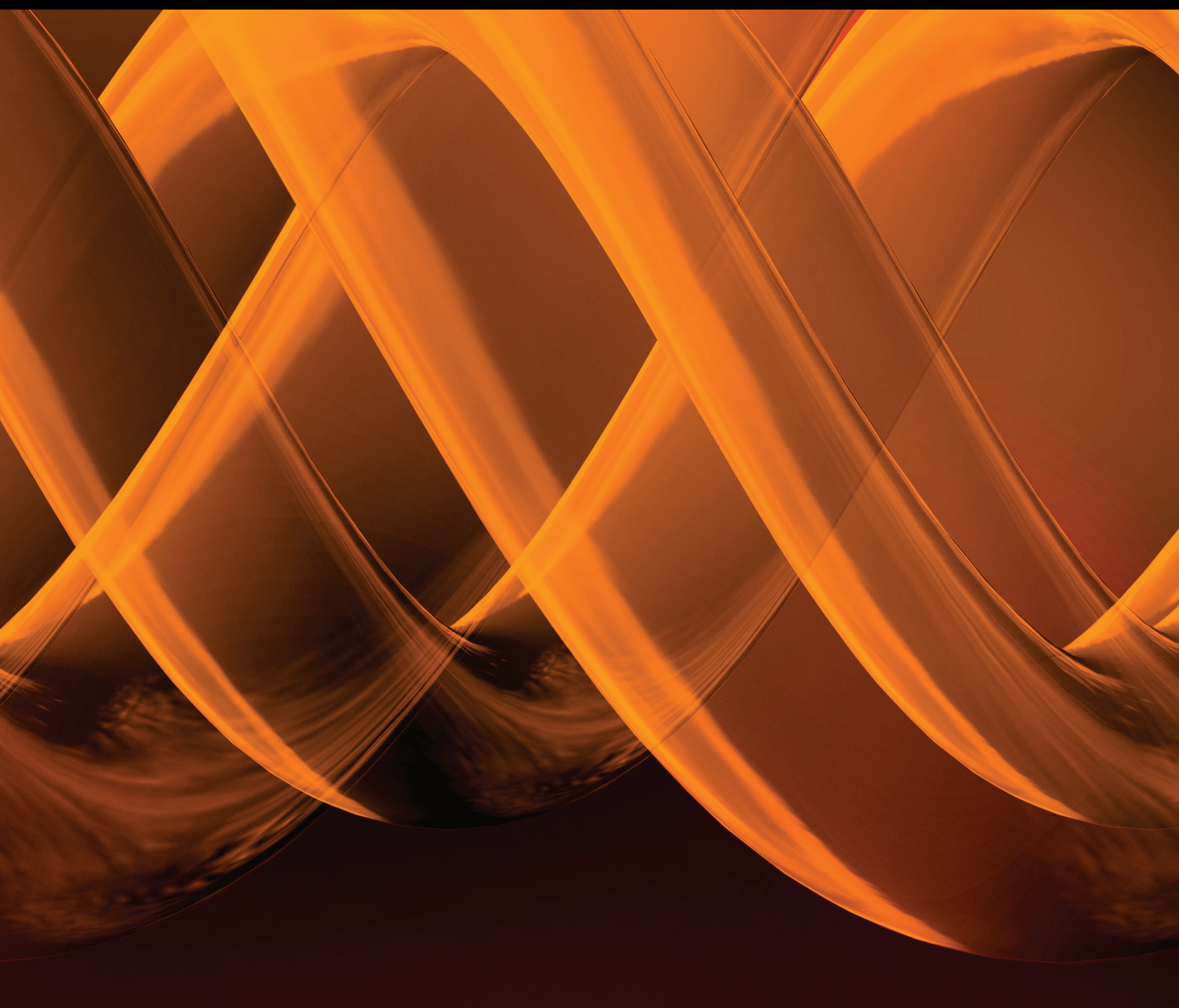
International Journal of Genomics

Recent Advances in High Throughput Sequencing Analysis

Lead Guest Editor: Yan Guo
Guest Editors: Leng Han and Quanhu Sheng



Recent Advances in High Throughput Sequencing Analysis

International Journal of Genomics

Recent Advances in High Throughput Sequencing Analysis

Lead Guest Editor: Yan Guo
Guest Editors: Leng Han and Quanhu Sheng

Editorial Board

Contents

Hindawi
International Journal of Genomics
Volume 2017, Article ID 2454780, 1 page
https://doi.org/10.1155/2017/2454780

Editorial

Recent Advances in High Throughput Sequencing Analysis

Yan Guo,[1] Leng Han,[2] and Quanhu Sheng[1]

[1]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37027, USA*
[2]*Department of Biochemistry and Molecular Biology, University of Texas Medical School, Houston, TX 77030, USA*

Correspondence should be addressed to Yan Guo; yan.guo@vanderbilt.edu

Received 22 March 2017; Accepted 22 March 2017; Published 19 June 2017

High-throughput genomic technology has enabled us to screen the entire genome and generate hypotheses at relatively low costs. One of the driving forces in high-throughput genomic technology is high-throughput sequencing (HTS). With its rapid development and affordability, HTS has quickly become the go-to choice for interrogating the entire genome. The analysis methodology development for HTS has been at the forefront of bioinformatics in recent years. Hundreds of tools and pipelines have been developed to aid researchers in interpreting HTS data. The aim of this special issue is to promote research and reflect the most recent advances in addressing HTS data analysis.

We received a total of 14 manuscripts and, through rigorous review, selected six for publication in this special issue. What follows is a brief summary of the six manuscripts:

(1) Title: "Chromosome 1 Sequence Analysis of C57BL/6J-Chr1^{KM} Mouse Strain." In this article, the authors studied the chromosome 1 sequence of the Chinese Kunming mouse and compared the sequence to three other mouse species.

(2) Title: "The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies." In this review article, the authors thoroughly examined the practicability of conducting high-throughput genomic assays, including HTS using formalin-fixed paraffin-embedded specimens.

(3) Title: "An Integrating Approach for Genome-Wide Screening of MicroRNA Polymorphisms Mediated Drug Response Alterations." In this study, the authors examined the relationship between polymorphisms and drug response in microRNA.

(4) Title: "Comparative Transcriptome Analysis Reveals Effects of Exogenous Hematin on Anthocyanin Biosynthesis during Strawberry Fruit Ripening." Through RNA sequencing, the authors examined the expression change of genes in strawberries that had been applied with exogenous hematin.

(5) Title: "Differential Gene Expression during Larval Metamorphic Development in the Pearl Oyster, *Pinctada fucata*, Based on Transcriptome Analysis." Through RNA sequencing, the authors studied changes in the gene expression pattern during the metamorphic development of a pearl oyster.

(6) Title: "RNA Sequencing of Formalin-Fixed, Paraffin-Embedded Specimens for Gene Expression Quantification and Data Mining." In this study, the authors examined the efficiency of two ribosomal RNA deletion kits: Ribo-Zero and RNase H.

Acknowledgments

We would like to extend our gratitude to all authors who contributed to this special issue and all reviewers that helped us select the highest quality manuscripts.

Yan Guo
Leng Han
Quanhu Sheng

Hindawi
International Journal of Genomics
Volume 2017, Article ID 1712530, 9 pages
https://doi.org/10.1155/2017/1712530

Research Article

Chromosome 1 Sequence Analysis of C57BL/6J-Chr1^{KM} Mouse Strain

Fuyi Xu,[1] Tianzhu Chao,[1] Yiyin Zhang,[1] Shixian Hu,[1] Yuxun Zhou,[1] Hongyan Xu,[2] Junhua Xiao,[1] and Kai Li[1]

[1]*College of Chemistry, Chemical Engineering, and Biotechnology, Donghua University, Shanghai, China*
[2]*Department of Biostatistics and Epidemiology, Medical College of Georgia, Augusta University, Augusta, GA, USA*

Correspondence should be addressed to Junhua Xiao; xiaojunhua@dhu.edu.cn and Kai Li; likai@dhu.edu.cn

Received 15 December 2016; Revised 9 February 2017; Accepted 15 February 2017; Published 9 April 2017

Academic Editor: Leng Han

The Chinese Kunming (KM) mouse is a widely used outbred mouse stock in China. However, its genetic structure remains unclear. In this study, we sequenced the genome of the C57BL/6J-Chr1^{KM} (B6-Chr1^{KM}) strain, the chromosome 1 (Chr 1) of which was derived from one KM mouse. With 36.6× average coverage of the entire genome, 0.48 million single nucleotide polymorphisms (SNPs) and 96,679 indels were detected on Chr 1 through comparison with reference strain C57BL/6J. Moreover, 46,590 of them were classified as novel mutations. Further functional annotation identified 155 genes harboring potentially functional variants, among which 27 genes have been associated with human diseases. We then performed sequence similarity and Bayesian concordance analysis using the SNPs identified on Chr 1 and their counterparts in three subspecies, *Mus musculus domesticus*, *M. m. musculus*, and *M. m. castaneus*. Both analyses suggested that the Chr 1 sequence of B6-Chr1^{KM} was predominantly derived from *M. m. domesticus* while 9.7% of the sequence was found to be from *M. m. musculus*. In conclusion, our analysis provided a detailed description of the genetic variations on Chr 1 of B6-Chr1^{KM} and a new perspective on the subspecies origin of KM mouse which can be used to guide further genetic studies with this mouse strain.

1. Introduction

The Chinese Kunming (KM) mouse colony, the largest outbred mouse stock maintained by commercial dealers nationwide in China, has been widely used in pharmaceutical and genetic studies [1]. Unlike other outbred mice, KM mouse has a complex evolutionary history. In 1944 during the World War II, Swiss mice were initially introduced into Kunming, Yunnan Province, China, from the Indian Haffkine Institute by Professor Feifan Tang via the Hump route with the help of the American Volunteer Group [2]. These mice were named KM mice after their initial location in China. Because most other mouse strains were lost and mouse facilities were damaged during the World War II, KM mouse became the only laboratory mouse available afterwards. They were gradually distributed throughout most of the country for medical studies. However, despite the importance of this outbred mouse, its underlying genetic structure remains unclear.

According to the Mouse Genome Informatics (http://www.informatics.jax.org/), over one thousand quantitative trait loci (QTLs) have been mapped on mouse chromosome 1 (hereafter referred to as Chr 1) including large amounts of QTLs related to metabolism disorder. However, very few candidate genes have been identified partly because of the large QTL intervals. In order to fine map the metabolism disorder QTLs on Chr 1 and identify the candidate genes, we established a population of Chr 1 substitution mouse strains, in which C57BL6/J (B6) was the host strain, and one KM mouse, five inbred strains, and twenty-four wild mice captured from various locations in China were selected as the Chr 1 donors [3]. In order to dissect the genetic structure and variations of this population and better severe further genetic studies, we have resequenced 18 strains

of this population including C57BL/6J-Chr1^{KM} (B6-Chr1^{KM}) with next-generation sequencing technology [4].

In this study, we analyzed the genome sequence data from B6-Chr1^{KM} strain and identified 0.48 million single nucleotide polymorphisms (SNPs) and 96,679 indels on Chr 1, of which 6.4% SNPs and 16.3% indels were considered to be novel. Functional annotation suggested that 474 variants had deleterious effect on gene functions. In addition, we explored the KM mouse genetic structure by performing sequence similarity and Bayesian concordance analysis (BCA) on Chr 1. Results suggested that KM mouse was predominately originated from *Mus musculus domesticus* and part of the sequence was from *M. m. musculus*.

2. Materials and Methods

2.1. Animals. B6 and KM mice were purchased from Shanghai SLAC Laboratory Animal Co., Ltd., China. One male KM mouse was mated with female B6 to produce hybrid F1, followed by 8 generations of backcrossing with B6 using marker-assisted selection, then brother × sister mating to create a B6-Chr1^{KM} Chr 1 substitution strain [3]. All mice were maintained under specific pathogen-free (SPF) conditions according to the People's Republic of China Laboratory Animal Regulations, and the study was conducted in accordance with the recommendations of and was approved by the Laboratory Animal Committee of Donghua University.

2.2. DNA Sequencing. B6-Chr1^{KM} genomic DNA was extracted from tail tissue of a male mouse using an AxyPrep™ Multisource Genomic DNA Miniprep Kit (Axygen, Hangzhou, China) according to the manufacturer's protocol.

Purified genomic DNA was sheared and size selected (300–500 bp). Paired-end sequencing (2 × 125 bp) was carried out with an Illumina HiSeq 2500 instrument (Illumina Inc., San Diego, CA, USA) on two lanes by WuXi AppTec (Shanghai, China) according to the manufacturer's protocol.

2.3. Read Alignment. Raw reads were filtered using NGS QC toolkit v2.3 [5] to remove reads containing more than 30% low-quality (Q20) bases. Filtered reads were aligned to the C57BL/6J reference genome (December 2011 release of the mouse reference genome (mm10) from Ensembl) using BWA (version 0.7.10-r789) with 12 threads [6]. The resulting SAM file was converted to a binary format and sorted with SAMtools v1.1 [7], followed by the marking of duplicate reads using picard-tools v1.119 (http://picard.sourceforge.net). To improve SNP and indel calling, indel realignment was conducted with Genome Analysis Toolkit (GATK v3.3) [8].

2.4. SNP/Indel Identification and Annotation. SNPs and indels were called using SAMtools mpileup and BCFtools call functions [7], with the '-uf' and '-cv' parameters, respectively. To identify a high-quality variant data set, variants were filtered using the BCFtools filter and VCFtools varFilter function [9]. The following parameters were used: for BCFtools filter, '-g 10 -G 3 -i 'QUAL>10 && MIN(MQ)>25 && MIN(DP)>6 && MAX(DP)<199 && (DP4[2]+DP4[3]) > 2', and for VCFtools varFilter, '-2 0'.

Ensembl Variant Effect Predictor tool (VEP, v78) [10] was used to characterize the SNPs and indels, and the algorithm SIFT was used to predict whether a missense variant would have a deleterious effect on a protein-coding gene.

2.5. Sequence Similarity Analysis. SNP information for WSB/EiJ (WSB), PWK/PhJ (PWK), and CAST/EiJ (CAST) was downloaded from the Mouse Genome Project (MGP) database of the Sanger Institute. The Chr 1 consensus sequence for each strain was constructed using the SAMtools consensus parameters. The repeat-masked B6-Chr1^{KM} Chr 1 sequence was divided into 1955 100 kb segments. The similarities of each segment with the corresponding segments in the WSB, CAST, and PWK were evaluated. Sliding window similarity analysis was also performed using 500 kb windows and 100 kb sliding intervals.

2.6. Phylogenetic Analysis. Phylogenetic analysis was conducted with the previously reported BCA method [11], with the *Rattus norvegicus* Chr 1 sequence (version rn5) downloaded from Ensembl used as the out-group. Briefly, consensus sequences from the WSB, PWK, and CAST strains were mapped to the alignment and gaps filled with Ns. Collinear segments were partitioned into 830 loci using a minimum description length algorithm with a default maximum cost.

2.7. Phylogenetic Tree Evaluation. Nexus files corresponding to the WSB-derived or PWK-derived regions were converted to FASTA files, and then a neighbor-joining phylogenetic tree was constructed using MEGA6 program [12]. Subsequently, 1000 bootstrap replicates were performed to generate branch support values.

3. Results

3.1. B6-Chr1^{KM} Genome Background. Chromosome substitution strains, also named as consomic strains, are designed to simplify the genome background and increase the power and speed of QTL mapping. The characteristic of consomic strain is that it only contains a single chromosome from the donor strain substituting the corresponding chromosome in the host strain. For B6-Chr1^{KM} consomic strain, Chr 1 sequence was derived from one KM mouse, while the genome background was from the B6 strain (Figure 1). In addition, sequences in the primary mouse reference assembly come from the same B6 strain. Therefore, our analysis of B6-Chr1^{KM} whole genome resequencing data only focused on Chr 1.

3.2. SNP and Indel Discovery. In this study, approximately one billion reads from the B6-Chr1^{KM} mouse strain were generated on two lanes of Illumina HiSeq 2500. A total of 78.65% of the reads were considered to be clean reads after quality control evaluation. Of them, more than 99% were aligned to the B6 mouse reference genome (mm10) using BWA with a mean genome-wide coverage of 36.6×.

A total of 479,956 SNPs and 96,679 indels were detected using SAMtools/BCFtools on Chr 1, in which 462,755 (96.42%) of the sites were homozygous. These variants were compared with variant calls from 36 key mouse strains from

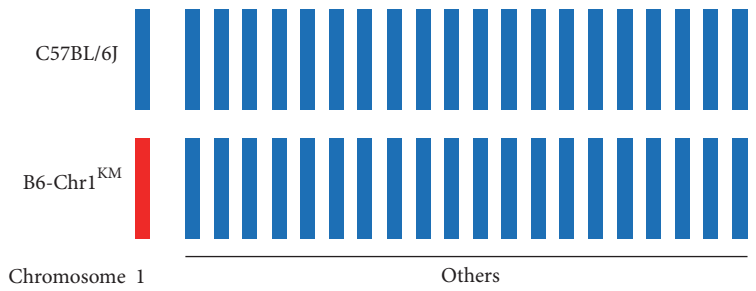


Figure 1: The characteristics of B6-Chr1^{KM} genome background. Blue bars represent B6 chromosome while the red represents KM mouse chromosome.

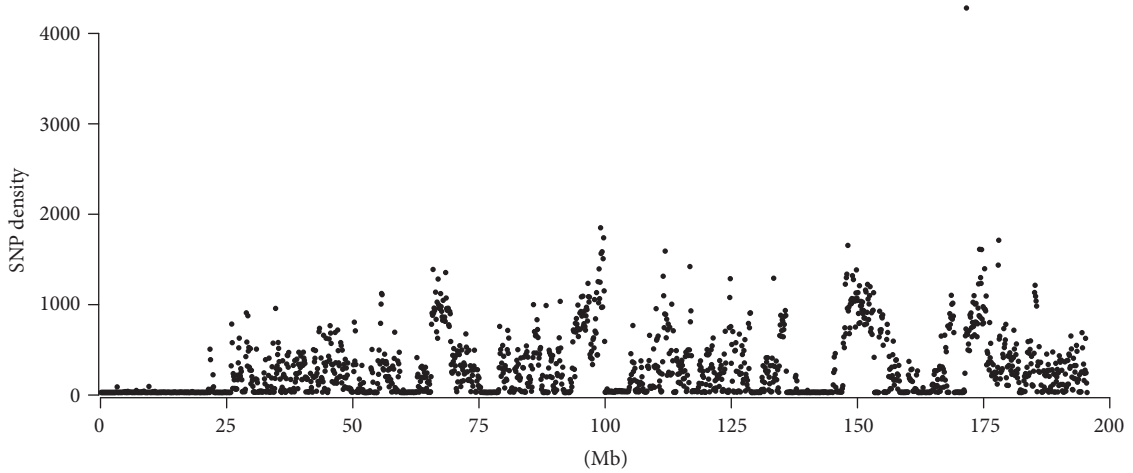


Figure 2: Distribution of SNP density on B6-Chr1^{KM} Chr 1. The SNP density is represented by the number of SNPs mapped within 100 kb physical intervals across Chr 1.

the Sanger Institute [13] as well as NCBI dbSNP142 variant data sets. This led to the identification of 449,089 SNPs (93.6%) as known, and the remaining 30,867 SNPs (6.4%) were classified as novel. For indels, 15,723 (16.3%) were classified as novel. In addition, we evaluated the variant calls using Sanger sequencing in our previous study which achieved high accuracy with 0.57% false positive and 0% false negative rate [4].

Next, we detected the distribution and density of SNPs over 100 kb window sizes. The observed average SNP density across the entire Chr 1 was 250 per 100 kb. However, different regions showed varying densities. For example, 29.5% of the Chr 1 sequence had an extremely low (0–5 SNPs per 100 kb) SNP density, while 9.1% had a high density (800 or more SNPs per 100 kb). The proximal region of Chr 1 was the longest region with a low SNP density encompassing nearly 25 Mb (Figure 2).

3.3. Functional Consequences of the SNPs and Indels. The putative consequences of SNPs and indels were cataloged using VEP from Ensembl (Table 1). The majority of the SNPs were located in intergenic (224,557, 18.7%) and intronic regions (575,013, 47.8%), and nearly 12% were classified as noncoding transcript variants. With regard to splice sites, 40 splice variants (including splice donor and splice acceptor variants) were found. The numbers of SNPs causing a premature stop codon or stop loss were 19 and 5, respectively. In addition, 2,378 (0.2%) missense variants were detected in 358 genes (one or more variants per gene). Among them, 380 variants (31.6%) from 113 genes were considered to have deleterious effects (SIFT < 0.05). Similar to the SNPs, the majority of indels were intronic (49.3%) and intergenic (17.1%) or within 5 kb upstream or downstream of a gene (16.9%). Only a small number of indels caused frameshift (22) and stop gain or loss (2). Among the novel variants, 7 caused a disruption of the translational reading frame; 10 were predicted as premature truncation of the protein due to gain or loss of stop codons; and 9 were located in splice donor regions. In addition, 104 novel missense variants from 20 genes had deleterious effects.

Next, we annotated these genes containing amino acid altering variants (SIFT < 0.05) and those with stop gain or loss, frameshift, and splice region variant genes with the Human-Mouse: Disease Connection database from Mouse Genome Informatics [14]. This analysis, which contained 155 genes, resulted in 27 genes associated with 49 different human disease-related phenotypes (Table 2), including macular degeneration, breast cancer, and immunodeficiency. Among these 27 disease genes, 9 have been investigated with mouse models, which had an in-depth phenotype information in different mouse genome background.

3.4. Sequence Similarity Analysis. The house mouse, *Mus musculus*, consists of three principal subspecies, with *M. m. domesticus* in Western Europe and the Middle East, *M. m.*

Table 1: Predictions of functional consequences of SNPs and indels.

Consequences	SNPs	Novel SNPs	Indels	Novel indels
splice_donor_variant	29	9	2	0
splice_acceptor_variant	11	0	4	0
stop_gained	19	8	1	1
frameshift_variant	0	0	22	7
stop_lost	5	1	1	0
start_lost	11	3	2	0
missense_variant	2378	486	—	0
inframe_insertion	0	0	28	2
inframe_deletion	0	0	26	2
splice_region_variant	1117	63	244	18
synonymous_variant	4238	281	0	0
stop_retained_variant	3	0	0	0
coding_sequence_variant	1	0	1	0
mature_miRNA_variant	4	2	2	0
5_prime_UTR_variant	1210	86	198	31
3_prime_UTR_variant	6563	484	1617	191
non_coding_transcript_exon_variant	11,955	640	2140	290
intron_variant	575,013	36,838	139,815	21,458
NMD_transcript_variant	42,052	2609	10,291	1372
non_coding_transcript_variant	143,110	8770	32,985	5190
upstream_gene_variant	96,888	8357	24,321	3992
downstream_gene_variant	93,184	5752	23,557	3390
intergenic_variant	224,557	13,198	48,568	8312

Consequences were predicted using Ensembl VEP and gene models from Ensembl version 76. Novel SNPs or indels are defined as variants that were not in MGP and dbSNP142 data sets.

musculus in Eastern Europe and Asia, and *M. m. castaneus* in Southeast Asia and India. Three genome sequences of the wild-derived inbred mouse strains, WSB, PWK, and CAST, which are broadly used to represent each of the subspecies, were selected for phylogenetic analysis. A Chr 1 consensus sequence was constructed for each strain using the SNP information from MGP. Because the simplest way to analyze phylogenetic divergence is by assessing sequence similarity, the Chr 1 sequence was separated into 1955 100 kb blocks and the similarities between each fragment and the corresponding sequences from WSB, PWK, and CAST were determined. The Chr 1 sequence was found to contain a large number of fragments with high sequence similarity to the corresponding sequence in WSB (Figure 3(a)), which is consistent with previous reports showing that KM mouse is derived from Swiss mice originated from the *M. m. domesticus* subspecies [1]. In addition, a bimodal distribution of blocks with two peaks of similarity was observed in a comparison of B6-Chr1^{KM} Chr 1 with PWK counterpart (Figure 3(a)). The first peak had only 99.05–99.1% sequence similarity to PWK, indicating the intersubspecies genome divergence of the Chr 1 sequence from *M. m. musculus*. The second peak had >99.7% sequence similarity to PWK (Figure 3(a)), indicating that the sequence of *M. m. musculus* introgressed into the KM mouse Chr 1. For the comparison of B6-Chr1^{KM} and CAST, we just observed one peak which suggested no signs of introgression of *M. m. castaneus* into the KM mouse Chr 1.

We next performed sliding window similarity analysis using 500 kb windows and 100 kb sliding intervals (Figure 3(b)). We found that 13.5% and 6.4% of the Chr 1 sequences had high similarity (>99.7%) with the corresponding sequences of PWK and CAST, respectively. The distal portion of the B6-Chr1^{KM} Chr 1 was found to have several regions that were highly similar to the corresponding regions of PWK with sharp boundaries between the regions of high and low similarity. However, we did not find any distinct boundaries between B6-Chr1^{KM} and CAST Chr 1 sequence.

3.5. Bayesian Concordance Analysis. To determine the extent of phylogenetic discordance in B6-Chr1^{KM} Chr 1, we assessed the discordance along Chr 1 by BCA. A total of 886 partitioned individual locus trees were used to estimate Bayesian concordance factors. In BCA, 87.7% of the loci supported a single KM/WSB topology with higher posterior probability, and 9.7% supported a single KM/PWK topology. None of the loci supported a KM/CAST topology, and the remaining 2.6% had a complicated topology (Figure 4(a)). Highly conserved genomic regions (Figure 3(b)) between the KM and PWK were almost found to have a relatively close topological relationship (Figure 4(a)). Furthermore, five loci with KM/WSB or KM/PWK topology were randomly

Table 2: List of human disease-associated genes with loss of function variants in B6-Chr1^{KM} Chr 1.

Gene	Ensembl ID	Variant type	Phenotype	OMIM ID
Col4a3	ENSMUSG00000079465	Frameshift	Alport syndrome, autosomal dominant	*104200*
			Alport syndrome, autosomal recessive	*203780*
			Hematuria, benign familial; BFH	*141200*
Fn1	ENSMUSG00000026193	Frameshift	Glomerulopathy with fibronectin deposits 2; GFND2	601894
			Plasma fibronectin deficiency	614101
Pde6d	ENSMUSG00000026239	Splice donor	Joubert syndrome 22; JBTS22	615665
Hmcn1	ENSMUSG00000066842	Frameshift	Macular degeneration, age-related, 1; ARMD1	603075
Cd244	ENSMUSG00000004709	Stop gain; splice donor	Rheumatoid arthritis; RA	180300
Rab3gap2	ENSMUSG00000039318	Splice acceptor, missense	Martsolf syndrome	212720
			Warburg micro syndrome 2; WARBM2	614225
Lamb3	ENSMUSG00000026639	Splice acceptor	Amelogenesis imperfecta, type IA; AI1A	104530
			Epidermolysis bullosa, junctional, Herlitz type	*226700*
			Epidermolysis bullosa, junctional, non-Herlitz type	*226650*
Dst	ENSMUSG00000026131	Missense	Epidermolysis bullosa simplex, autosomal recessive 2; EBSB2	615425
			Neuropathy, hereditary sensory and autonomic, type VI; HSAN6	*614653*
Ercc5	ENSMUSG00000026048	Missense	Xeroderma pigmentosum, complementation group G; XPG	278780
Casp8	ENSMUSG00000026029	Missense	CASPase 8 deficiency	*607271*
			Dermatitis, atopic	*603165*
Tmem237	ENSMUSG00000038079	Missense	Joubert syndrome 1; JBTS1	213300
			Joubert syndrome 14; JBTS14	614424
Bard1	ENSMUSG00000026196	Missense	Breast cancer	*114480*
Bcs1l	ENSMUSG00000026172	Missense	Bjornstad syndrome; BJS	262000
			Gracile syndrome	*603358*
			Leigh syndrome; LS	256000
			Mitochondrial complex III deficiency, nuclear type 1; MC3DN1	124000
Obsl1	ENSMUSG00000026211	Missense	Three M syndrome 2; 3 M2	612921
Tm4sf20	ENSMUSG00000026149	Missense	Specific language impairment 5; SLI5	615432
Dis3l2	ENSMUSG00000053333	Missense	Perlman syndrome; PRLMNS	267000
Chrng	ENSMUSG00000026253	Missense	Multiple pterygium syndrome, Escobar variant; EVMPS	265000
			Multiple pterygium syndrome, lethal type; LMPS	253290
Ugt1a1	ENSMUSG00000089960	Missense	Crigler-Najjar syndrome, type I	218800
			Crigler-Najjar syndrome, type II	606785
			Gilbert syndrome	143500
			Hyperbilirubinemia, transient familial neonatal; HBLRTFN	237900
Steap3	ENSMUSG00000026389	Missense	Anemia, hypochromic microcytic, with iron overload 2; AHMIO2	615234
Ube2t	ENSMUSG00000026429	Missense	Fanconi anemia, complementation group T; FANCT	616435
Ppox	ENSMUSG00000062729	Missense	Porphyria variegata	*176200*
Ackr1	ENSMUSG00000037872	Missense	Malaria, susceptibility to	611162
Spta1	ENSMUSG00000026532	Missense	Elliptocytosis 2; EL2	130600
			Pyropoikilocytosis, hereditary; HPP	266140
			Spherocytosis, type 3; SPH3	*270970*
Ephx1	ENSMUSG00000038776	Missense	Epoxide hydrolase 1, microsomal; EPHX1	132810
			Hypercholanemia, familial; FHCA	607748
			Preeclampsia/eclampsia 1; PEE1	189800
Rd3	ENSMUSG00000049353	Missense	Leber congenital amaurosis 12; LCA12	*610612*

Table 2: Continued.

Gene	Ensembl ID	Variant type	Phenotype	OMIM ID
Cd46	ENSMUSG00000016493	Missense	Hemolytic uremic syndrome, atypical, susceptibility to, 2; AHUS2	612922
			Immunodeficiency, common variable, 2; CVID2	240500
Cr2	ENSMUSG00000026616	Missense	Immunodeficiency, common variable, 7; CVID7	614699
			Systemic lupus erythematosus, susceptibility to, 9; SLEB9	610927

OMIM: online Mendelian inheritance in man. Numbers in italic in OMIM ID column indicate that these diseases have mouse models. Human disease-related phenotypes come from "Human-Mouse: Disease Connection" database (http://www.informatics.jax.org/humanDisease.shtml) in Mouse Genome Informatics website.

selected, and the phylogenetic trees were confirmed by Mega software (Figure 4(b)).

4. Discussion

Because the KM mouse is used regularly in pharmaceutical and genetic studies, its detailed genetic structure is of great value to the research community. In this study, we sequenced the genome of a male B6-Chr1^{KM} mouse, in which Chr 1 was derived from one KM mouse. The detailed sequence analysis would provide new insights into the application of B6-Chr1^{KM} in biomedical research.

In this study, we identified 479,956 SNPs and 96,679 indels on Chr 1, of which 8.1% did not exist in the MGP and dbSNP142 data sets, indicating that these variants were unique to the B6-Chr1^{KM} mice. Therefore, these variants can be used as unique genetic markers for the genetic quality control of KM mouse. As the most common types of genetic variants, SNPs and indels have been increasingly recognized as having a wide range of effects on gene functions. Among the variants identified on Chr 1, most were located within intergenic or intronic regions. However, we also identified 474 functional variants (missense variant with SIFT < 0.05, stop gain or loss variant, frameshift variant, and splice donor or acceptor variant) which influenced 155 genes. Additionally, several genes have been identified to be associated with human diseases, making them interesting candidates for further functional studies using KM mouse or our newly build B6-Chr1^{KM} strain. For example, Rd3, which is associated with retinal degeneration, was identified as a missense substitution (A->T) with significant deleterious effects ($p = 0.02$). Previous studies have shown that mice with a homozygous mutation in Rd3 exhibit retinal degeneration at three weeks after birth [15]. We also identified a splice acceptor variant in Lamb3 gene, which is associated with blistering of the skin. The mouse models with homozygous Lamb3 628 G->A showed blistering and erosions after birth [16].

Since KM mouse is originated from Swiss mice, it has been speculated to be contaminated with *M. m. castaneus*. In 1991, the morphological characteristics and isozyme polymorphisms of KM and Swiss mice were evaluated, revealing the presence of distinct genetic differences between them [17]. Comparison of KM mouse with wild mice of *M. m. castaneus* captured in Kunming has revealed that the former is more closely related to *M. m. domesticus* than to *M. m. castaneus*. Conversely, contamination of KM mouse by *M. m. castaneus* has been previously demonstrated using the isozyme test [18]. In 2003, the results of a study involving the detection of isozyme polymorphisms also supported the grouping of KM and Swiss mice with *M. m. domesticus* and not with *M. m. musculus* or *M. m. castaneus* [2]. However, it has not yet been confirmed whether KM mouse contains part of the genome of *M. m. musculus* or *M. m. castaneus*. Therefore, high resolution studies of Chr 1 of KM mouse by next-generation sequencing may clarify whether these mice were originated from Swiss mice and/or other mice. Our sequence similarity analysis provided substantial evidence that KM mouse was derived from *M. m. domesticus*, which means that Swiss mice were their ancestor. Both 100 kb blocks and sliding window similarity analysis demonstrated that the Chr 1 of KM mouse was largely composed of *M. m. domesticus* sequences with the rest may derive from *M. m. musculus* or *M. m. castaneus*. Therefore, further analysis is needed to determine the proportion of each subspecies contribution to the Chr 1 of KM mouse.

With the increasing number of whole genome data sets, the reconstruction of phylogenetic trees at a genomic scale has become feasible. Exploration of these large data sets has revealed that there may be discordance among the topologies in different genomic regions [19, 20]. Although these differences may be caused by incorrect estimations of gene genealogies, incongruent gene trees can also be attributed to the differing evolutionary histories of different genomic regions, especially for close species or subspecies. Traditionally, there are two types of phylogenetic analysis methods, the consensus method and the total evidence method. Both methods barely quantify the topological discordance across the entire genome. Recently, BCA, which is an improvement upon the consensus method, has been used to statistically quantify the discordance, as well as to generate phylogenetic trees [21]. A few studies using BCA have demonstrated its great potential for the reconstruction of phylogenic trees of mouse subspecies [11, 13, 22]. These studies indicate that BCA is a suitable method to quantify the proportions of Chr 1 sequence in B6-Chr1^{KM} derived from the different subspecies. Through BCA, we found approximately that 90% and 10% of the sequences of Chr 1 were derived from *M. m. domesticus* and *M. m. musculus*, respectively. Although the sequence similarity analysis revealed that there were some regions which had higher sequence similarity with CAST, we did not observed the same results in the BCA. Therefore, we cannot make the

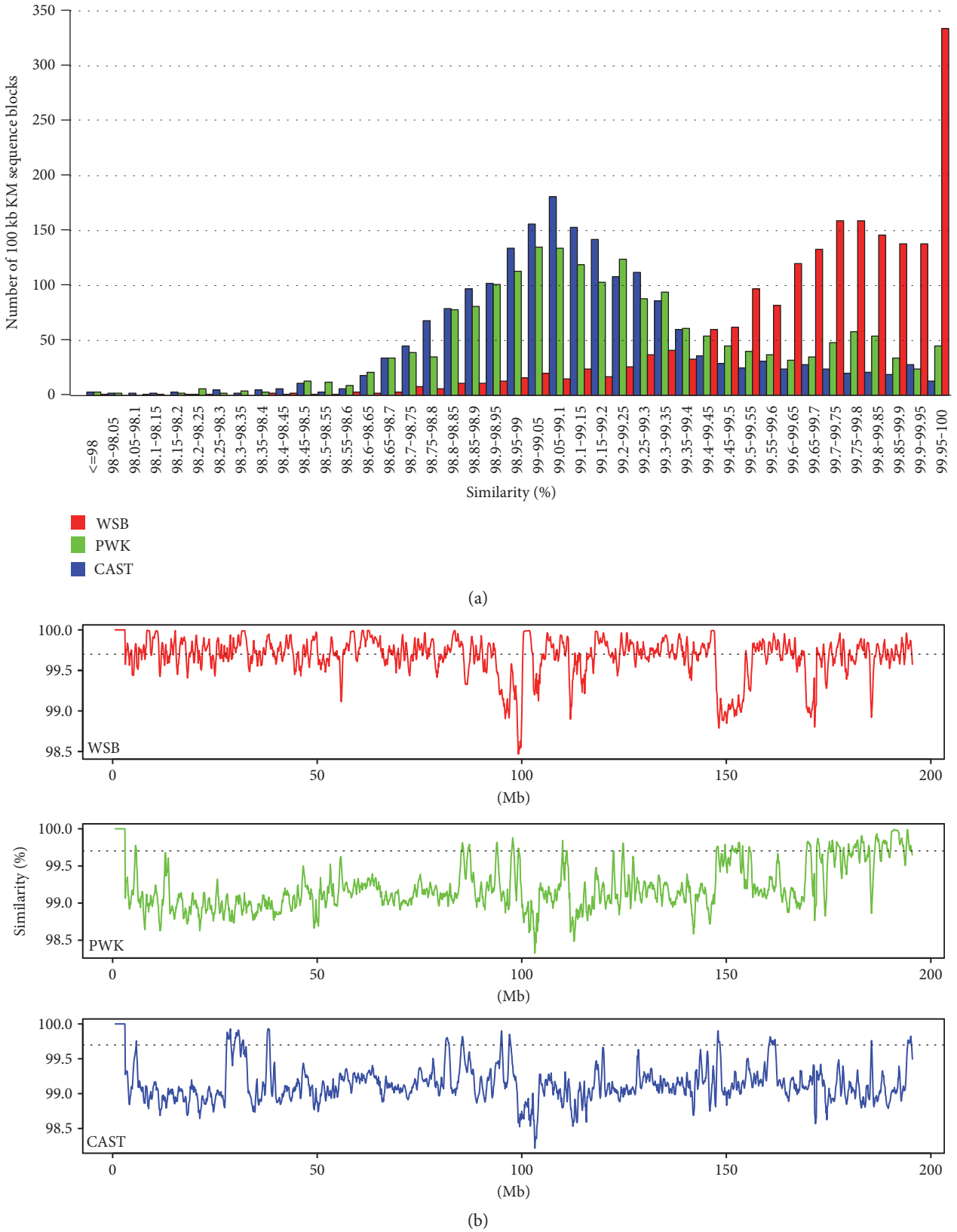


Figure 3: Sequence similarity between B6-Chr1^{KM} and WSB, PWK, and CAST Chr 1. (a) Distribution of the numbers of 100 kb blocks of the B6-Chr1^{KM} Chr 1 with sequence similarities (%) to the corresponding blocks of the WSB, PWK, and CAST Chr 1. (b) Sliding window analysis of the similarities of Chr 1 sequences between B6-Chr1^{KM} and WSB, CAST, or PWK. The B6-Chr1^{KM} Chr 1 sequence was compared using 500 kb windows and 100 kb sliding intervals. The horizontal line indicates the level of 99.7% sequence similarity.

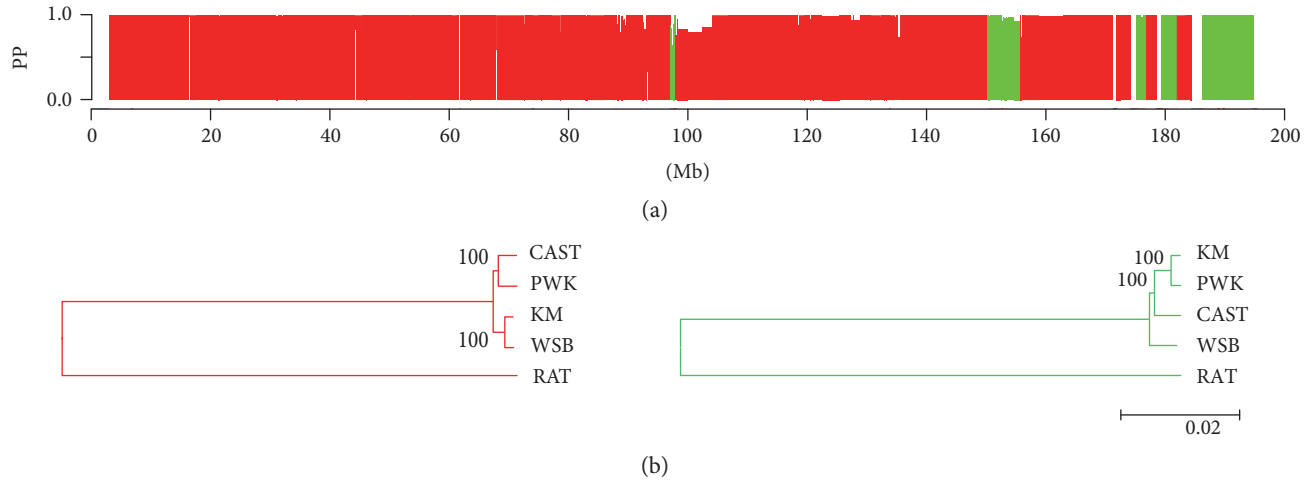


Figure 4: Phylogenetic analysis of B6-Chr1^{KM} Chr 1. (a) Fine-scale phylogenetic discordance of B6-Chr1^{KM} Chr 1 (PP indicates posterior probability). Red represents WSB, green indicates PWK, and white represents unknown. (b) Phylogenetic tree of the WSB-derived or PWK-derived sequences of B6-Chr1^{KM} and the wild-derived inbred mouse strain sequences. A neighbor-joining tree was generated using MEGA6 software. Red and green indicate regions supporting a single topology for KM/WSB and KM/PWK, respectively, which are both associated with a high posterior probability, as determined by BCA.

conclusion that some of Chr 1 sequence of B6-Chr1^{KM} came from CAST which represent *M. m. castaneus*. While for PWK, highly conserved genomic regions (Figure 3(b)) with KM aligned well with the BCA results (Figure 4(a)). Thus, from both analyses, we can make the conclusion that Chinese KM mouse has a mosaic genome structure with sequences predominately derived from *M. m. domesticus* and with at least some of the remaining sequences derived from *M. m. musculus*.

In summary, we presented the analysis of a high-quality genome sequence of the B6-Chr1^{KM}. These data allow better understanding of the structure and origin of the genetic variations in the B6-Chr1^{KM} mouse strain, which provides insights into the utility of this mouse strain and the KM outbred stock for further biomedical research and the study of complex diseases.

Data Access

All raw reads were submitted to NCBI Sequence Read Archive under the Accession no. SRR2954707 associated with BioProject Accession no. PRJNA298468 and BioSample Accession no. SAMN04159475.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Fuyi Xu and Tianzhu Chao contributed equally to this work.

Acknowledgments

This work was supported by the Key Project of Science & Technology Commission of Shanghai Municipality (no. 13140900300), the National Science Foundation of China (no. 31171199), the Fundamental Research Funds for the Central Universities (no. 2232013A3-06), and the DHU Distinguished Young Professor Program (B201308).

References

[1] X. Zhang, Z. Zhu, Z. Huang, P. Tan, and R. Z. Ma, "Microsatellite genotyping for four expected inbred mouse strains from KM mice," *Journal of Genetics and Genomics*, vol. 34, no. 3, pp. 214–222, 2007.

[2] B. Yue, S. Liu, D. Liu et al., "Comparative studies on the genetic biological markers of five closed colonies of Kunming mice," *Laboratory Animal Science and Management*, vol. 20, no. S1, pp. 58–62, 2003.

[3] J. Xiao, Y. Liang, K. Li et al., "A novel strategy for genetic dissection of complex traits: the population of specific chromosome substitution strains from laboratory and wild mice," *Mammalian Genome*, vol. 21, no. 7-8, pp. 370–376, 2010.

[4] F. Xu, T. Chao, Y. Liang et al., "Genome sequencing of chromosome 1 substitution lines derived from Chinese wild mice revealed a unique resource for genetic studies of complex traits," *G3 (Bethesda)*, vol. 6, no. 11, pp. 3571–3580, 2016.

[5] R. K. Patel and M. Jain, "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data," *PloS One*, vol. 7, no. 2, article e30619, 2012.

[6] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[7] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[8] A. McKenna, M. Hanna, E. Banks et al., "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[9] P. Danecek, A. Auton, G. Abecasis et al., "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[10] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–2070, 2010.

[11] M. A. White, C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur, "Fine-scale phylogenetic discordance across the house mouse genome," *PLoS Genetics*, vol. 5, no. 11, article e1000729, 2009.

[12] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[13] T. M. Keane, L. Goodstadt, P. Danecek et al., "Mouse genomic variation and its effect on phenotypes and gene regulation," *Nature*, vol. 477, no. 7364, pp. 289–294, 2011.

[14] J. A. Blake, J. T. Eppig, J. A. Kadin et al., "Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse," *Nucleic Acids Research*, vol. 45, no. D1, pp. D723–D729, 2017.

[15] J. S. Friedman, B. Chang, C. Kannabiran et al., "Premature truncation of a novel protein, RD3, exhibiting subnuclear localization is associated with retinal degeneration," *American Journal of Human Genetics*, vol. 79, no. 6, pp. 1059–1070, 2006.

[16] J. Hammersen, J. Hou, S. Wunsche, S. Brenner, T. Winkler, and H. Schneider, "A new mouse model of junctional epidermolysis bullosa: the LAMB3 628G>A knockin mouse," *The Journal of Investigative Dermatology*, vol. 135, no. 3, pp. 921–924, 2015.

[17] S. Shi, H. Wang, B. Cui et al., "Study on genetic variants of Chinese KM mouse subcolonies," *Chinese Journal of Laboratory Animal Science*, vol. 1, no. 1, pp. 29–36, 1991.

[18] G. Zhao, S. Bao, D. Zhang, R. Zhang, and M. Jin, "Probing into the course of formation of genetic character of KM mouse," *Shanghai Laboratory Animal Science*, vol. 14, no. 1, pp. 1–4, 1994.

[19] D. A. Pollard, V. N. Iyer, A. M. Moses, and M. B. Eisen, "Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting," *PLoS Genetics*, vol. 2, no. 10, article e173, 2006.

[20] G. Giribet, G. D. Edgecombe, and W. C. Wheeler, "Arthropod phylogeny based on eight molecular loci and morphology," *Nature*, vol. 413, no. 6852, pp. 157–161, 2001.

[21] C. Ane, B. Larget, D. A. Baum, S. D. Smith, and A. Rokas, "Bayesian estimation of concordance among gene trees," *Molecular Biology and Evolution*, vol. 24, no. 2, pp. 412–426, 2007.

[22] T. Takada, T. Ebata, H. Noguchi et al., "The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains," *Genome Research*, vol. 23, no. 8, pp. 1329–1338, 2013.

Hindawi
International Journal of Genomics
Volume 2017, Article ID 1674827, 7 pages
https://doi.org/10.1155/2017/1674827

Research Article

An Integrating Approach for Genome-Wide Screening of MicroRNA Polymorphisms Mediated Drug Response Alterations

Xianyue Wang,[1] Hong Jiang,[1] Wei Wu,[1] Rongxin Zhang,[1] Lingxiang Wu,[1] Huan Chen,[1] Pengping Li,[1] Yumin Nie,[1] Jiaofang Shao,[1] Yan Li,[1] Xue Lin,[1] Sali Lv,[1,2,3] Qh Wang,[1,2,3] and Jie Hu[1]

[1]*Department of Bioinformatics, Nanjing Medical University, Nanjing 210029, China*
[2]*Key Laboratory of Human Functional Genomics of Jiangsu Province, Nanjing Medical University, Nanjing, China*
[3]*Collaborative Innovation Center for Cardiovascular Disease, Nanjing Medical University, Nanjing, China*

Correspondence should be addressed to Qh Wang; wangqh@njmu.edu.cn and Jie Hu; hujie@njmu.edu.cn

Received 20 October 2016; Accepted 20 December 2016; Published 5 April 2017

Academic Editor: Quanhu Sheng

MicroRNAs (miRNAs) are a class of evolutionarily conserved small noncoding RNAs, ~22 nt in length, and found in diverse organisms and play important roles in the regulation of mRNA translation and degradation. It was shown that miRNAs were involved in many key biological processes through regulating the expression of targets. Genetic polymorphisms in miRNA target sites may alter miRNA regulation and therefore result in the alterations of the drug targets. Recent studies have demonstrated that SNPs in miRNA target sites can affect drug efficiency. However, there are still a large number of specific genetic variants related to drug efficiency that are yet to be discovered. We integrated large scale of genetic variations, drug targets, gene interaction networks, biological pathways, and seeds region of miRNA to identify miRNA polymorphisms affecting drug response. In addition, harnessing the abundant high quality biological network/pathways, we evaluated the cascade distribution of tarSNP impacts. We showed that the predictions can uncover most of the known experimentally supported cases as well as provide informative candidates complementary to existing methods/tools. Although there are several existing databases predicting the gain or loss of targeting function of miRNA mediated by SNPs, such as PolymiRTS, miRNASNP, MicroSNiPer, and MirSNP, none of them evaluated the influences of tarSNPs on drug response alterations. We developed a user-friendly online database of this approach named Mir2Drug.

1. Introduction

MicroRNAs (miRNAs) are a class of evolutionarily conserved small noncoding RNAs, 19~25 nt in length, and found in diverse organisms [1]. MiRNAs play important roles by binding to the 3′-untranslated region (3′ UTR) of the target mRNA, causing the reduction of its abundance and translational efficiency [2]. It has been shown that miRNAs are involved in many biological processes of human complex diseases through regulating gene expression [3]. Many of the target genes of miRNAs are drug targets, and the genes involved in drug disposition may also be regulated by miRNAs [4]. Therefore, genetic polymorphisms in miRNA target sites may alter the miRNA regulation of these drug-related genes and result in the differential expression of the drug-related protein, which can in turn influence the drug response.

Recently, as the number of drug-related SNPs in miRNA target sites are rapidly increasing, the importance of SNPs positioned in the 3′ UTR regions is becoming evident [5]. Several recent studies have demonstrated that the single nucleotide polymorphisms in miRNA target sites can affect drug efficiency. Mishra et al. showed that a functional SNP presents in 3′ UTR of dihydrofolate reductase, an important drug target, and the SNP interferes with the miR-24 miRNA function and leads to DHFR overexpression and methotrexate resistance [6]. Wynendaele et al. demonstrated that an SNP created an illegitimate miRNA target site within the 3′ UTR of MDM4, which affected ovarian cancer progression and chemo sensitivity [7]. Boni et al. also demonstrated that several SNPs had a significant association with clinical outcome of ovarian cancer patients treated with the 5-FU and CPT-11 combination [8]. Polymorphisms in the miRNA

target sites are emerging as powerful tools to elucidate the underlying mechanisms of different responses to treatments in patients. In addition, several other studies indicate that miRNAs can affect drug sensitivity and resistance in cancer chemo therapy [9]. Based on the above-mentioned experimental evidence, we have reason to believe that SNP in miRNA target sites affect drug response may be common, and there are still a large number of these specific genetic variants that have yet to be discovered. However, experimental approaches to identify these SNP that can affect drug response are labor-intensive and time-consuming. To address these challenges, computational and analytical tools can be developed to provide the successful design of biological experiments and interpretation of the results. Characterization of SNP in miRNA target sites in drug response helps predict patients' responses to drug treatments, guides rational drug use, and improves drug safety and efficacy.

Several studies have developed databases or tools that can predict SNPs reside in miRNA target sites. Bao et al. developed a database that collected naturally occurring DNA variations in putative miRNA target sites; this database integrates sequence polymorphism, phenotype, and expression microarray data [10]. Hiard et al. developed a database that compiled DNA sequence polymorphisms that are predicted to perturb miRNA mediated gene regulation. The database also includes the inclusion of copy number variants and eQTL information that affect miRNA precursors as well as genes encoding components of the silencing machinery [11]. Barenboim et al. developed a web-based tool that predicted the impact of an SNP on putative miRNA targets [12]. This application interrogates the 3′-untranslated region and predicts if an SNP within the target site will disrupt/eliminate or enhance/create a miRNA binding site. Hariharan et al. analyzed SNPs in and around predicted miRNA target sites; polymorphisms within 200 nucleotides that could alter miRNA regulation were annotated [13]. However, all existing tools above did not concern the correlation between polymorphisms in miRNA target sites and drug response. More specifically, in addition to variation located in the target site, variation near the target site has been identified as another crucial factor that can influence an individual's response to drugs. For example, an SNP is located 14 bp downstream of the miR-24 target site in DHFR 3′ UTR, does not directly fall within the miRNA target set, and resulted in DHFR overexpression and MTX resistance [6]. Instead, in this study, we propose to analyze all the SNPs near the target sites in 3′ UTR of all mRNA, to have a general overview of the SNPs' regulatory effect on the drug response.

We have developed a user-friendly online database, Mir2Drug (publicly accessible at http://bioinfo.njmu.edu.cn/Mir2Drug), which can help researchers to identify the SNPs, names as tarSNPs, which can affect drug response that does not only reside in target sites but also near target sites. We also identified those SNPs that indirectly affect drug response through protein-protein interaction (PPI) network and pathway. We show that the database's predictions can uncover most of the known experimentally supported cases and provide better performance than other existing databases and tools. Mir2Drug can be used to predict drug response to therapy and is useful for explaining the differences in drug response, discovery, and characterization of novel predictive and prognostic biomarkers.

2. Materials and Methods

2.1. Identifying SNPs Affected miRNA Regulation (tarSNPs). Mature miRNA sequences were derived from the miRBase, release 21 [14]. SNPs that are located in the 3′ UTRs of all known genes and 3′ UTRs sequences were retrieved from UCSC Genome browser (dbSNP build 147 and NCBI build 38) [15]. For each SNP, we extracted the 30 bp flanking sequence of both upstream and downstream of the SNP in the 3′ UTR region of genes. Then, we assessed whether the two alleles of an SNP lead to different miRNA target sites based on the 61 bp DNA sequence, using the program PITA [16] with default parameters. PITA predicts potential miRNA targets using an estimated free energy. The SNPs changing the free energy between miRNA and DNA sequence were defined as tarSNPs.

The degree of binding is quantified using the PITA score change. Mathematically, the binding degree is described as

$$\text{diff} = T_{\text{wt}} - T_{\text{var}}, \tag{1}$$

where T_{wt} represents the score of miRNA binding to the wild-type 61 bp DNA sequences using the program PITA and T_{var} represents the score of miRNA binding to the variant-type sequences. diff represents the degree of miRNA regulation change from wild-type allele to the variant-type; a positive diff represents a strengthen miRNA regulation ability from wild-type allele to the variant-type; on the contrary, a negative diff represents a weaken miRNA regulation ability. According to the binding of miRNA to the mRNA 3′ UTR, we assigned the potential tarSNPs to one of the four classes: "complete gain," the mRNA acquires a new target site through the wild-type SNP into variant-type SNP; "complete loss," mRNA loses a predicted target site through the wild-type SNP into variant-type SNP; "partial gain," mRNA acquires more stable target site than without the SNP; "partial loss," mRNA target site turns into instable target site with the SNP. For the scenario of multiple tarSNPs identified in a single patient, we utilize the normalized binding energy differences for prioritizing the tarSNPs, which is described as

$$\frac{|\text{diff}|}{\max\left(\left|T_{\text{wt}}\right|, \left|T_{\text{var}}\right|\right)}. \tag{2}$$

2.2. Mapping Predicted tarSNPs for Direct Drug Target. We have already got all genes that contained at least one tarSNP. Then, we mapped the predicted genes to drug target and extracted the genes by drug target and by containing tarSNP, we got all targets that at least one tarSNP is physically located in it. These drug target genes that contained tarSNP defined direct drug target. All known associations of drug and targets were downloaded from the DrugBank database (DrugBank 5.0) [17]. We set all DrugBank targets containing tarSNP as the direct drug targets.

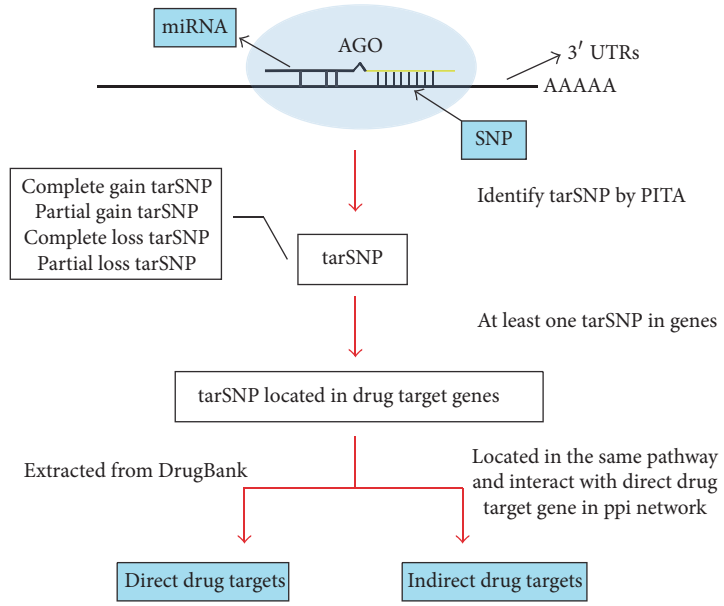


Figure 1: The workflow of the Mir2Drug database that identified SNPs affecting drug response.

2.3. Network Integrating for Prediction of Indirect Drug Target. We extract all the genes that interact with direct drug target genes in PPI network and defined them as a PPIN-indirect drug target. Totally, 4234 PPIN-indirect drug target genes were identified. The human PPI data was derived from the HPRD database [18]. We also get all the genes that are located in the same pathway with direct drug target genes, we defined them as pathway-indirect drug target, we get 2124 pathway-indirect drug target genes in the metabolic pathway and 5424 in the nonmetabolic pathway, and we computed the length of shortest paths for each pathway-indirect drug target genes and drug target genes. The metabolic and nonmetabolic pathways were downloaded from the Reactome database [19]. To assist the identification of tarSNPs that can indirectly affect drug response, we mapped the predicted tarSNPs to indirect drug target and extracted the genes by indirect drug target and by containing tarSNP.

2.4. Mir2Drug Database Construction. All useful results and information in this work were organized into a set of relational MySQL tables for fast access. HTML, CSS, JavaScript, and PHP were used to construct the online website Mir2Drug which performs multiple browse and search functions running on Apache web server (http://www.apache.org).

3. Results

3.1. Functional miRNA Polymorphisms Widely Distributed on Drug Targets. Unlike other tools (PolymiRTS [10, 20], MicroSNiPer [12], MirSNP [21], and miRNASNP [22]), we considered that both SNPs reside and near the target site can affect the interaction of miRNA to the target site (tarSNPs). Therefore, we also include all genes that are located in the same pathway with direct drug target genes and that interact with direct drug target genes in PPI network as target genes (Figure 1). At least, we got direct 3408 drug target genes and 9143 indirect drug target genes; we got all targets that at least one tarSNP is located in it. These genes are candidate drug response genes underlying tarSNP regulation.

For a SNP in the miRNA target, if it gives rise to more free energy changed, more instability of the structure of the miRNA/mRNA interaction occurs. In addition, 5188 pairs of experimentally confirmed miRNA/mRNA interaction downloaded from miRTarBase [28]. And the average free energy of the structures on miRNA/mRNA interaction is -4.58 ± 3.60 kcal/mol. The free energy change of the structures is a relatively high level ranging from 1 to 33.71. About 50% of energy changes are >2.5 kcal/mol, which may affect the stability of the structure significantly. In our results, a set of 833134 out of 973671 SNPs were identified as tarSNPs on 10021 drug targets. The tarSNPs have been generally fallen into four classes based on our methods. TarSNPs of class "complete gain" and "partial gain" may cause a gain of miRNA function and downregulation of the target protein; if the miRNA target protein is a drug target, its decreased level will result in drug sensitivity and vice versa: tarSNPs of class "complete loss" and "partial loss" may cause a loss of miRNA regulation; it will cause upregulation of the target protein, resulting in drug resistance. Both direct drug targets and indirect drug targets are featured in the database.

3.2. SNPs in 3′ UTR of Drug Target Affect Drug Efficiency Induced by Gain or Loss Function of miRNA Targeting. Some drug target genes showed a significant downregulation or upregulation in the mRNA level in drug-treated cells comparing to the nontreated cells. We supposed that drug treatment may cause variation in the drug target genes' 3′ UTR to change the stability of the structure of miRNA/mRNA interaction. Some research showed that BTG1 was upregulated in response to treatment with tomato leaf extract (TLE) [23],

Table 1: Drugs and their predicted differentially expressed targets in Mir2Drug database.

Drug	SNP ID	Gene symbol	miRNA	Wild free energy	Mutation free energy	$\Delta\Delta G$ (kcal/mol)	Effect class	Literature
TLE	rs764927448	BTG1	hsa-miR-1255b-5p	−0.26	0.8	−1.06	Partial loss	[23]
Recombinant bromelain	rs764927448	BTG1	hsa-miR-1255b-5p	−0.26	0.8	−1.06	Partial loss	[24]
IGF-I	rs764927448	BTG1	hsa-miR-1255b-5p	−0.26	0.8	−1.06	Partial loss	[25]
Anti-HER2/neu trastuzumab antibody	rs746218880	WT1	hsa-miR-1193	None	−8.11	8.11	Complete gain	[26]
Norlichexanthone and anomalin A	rs72552389	PIM1	hsa-miR-7114-5p	None	−11.12	11.12	Complete gain	[27]

recombinant bromelain [24], and insulin-like growth factor I (IGF-I) [25] in the MCF-7 breast cancer cell line, which contained a SNP (rs764927448) in BTG1 by query of the Cancer Cell Line Encyclopedia (CCLE). Moreover, we found that BTG1 loss is regulated by has-miR-1255b-5p through the wild-type SNP into variant-type SNP in Mir2Drug database (Table 1). Conversely, Tuna et al. showed that HER2/neu engages Akt to increase WT1 expression to stimulate S-phase proliferation and inhibit apoptosis in breast cancer cells, and then inhibition of HER2/neu with the anti-HER2/neu trastuzumab antibody decreased WT1 protein levels in HER2/neu-overexpressing BT-474 cells [26]. Ebada et al. showed that in vitro IC50 values of Norlichexanthone and anomalin A inhibited PIM1 in A2780 ovary cancer cell line [27]. We also found an SNP (rs746218880) in WT1 3′ UTR and an SNP (rs72552389) in PIM1 3′ UTR, which create miRNA/mRNA interaction in Mir2Drug database (Table 1).

3.3. Harnessing the Molecular Networks Mir2Drug Increased the Efficiency of Identifying Novel tarSNPs as well as Recapitulating Known tarSNPs. To demonstrate the effectiveness of our optimized method of recapitulating known cases where SNPs affected the miRNA binding, we manually compiled nine experimentally confirmed tarSNPs using text mining (Table 2). Although there are several existing databases predicting the gain or loss of targeting function of miRNA mediated by SNPs, such as PolymiRTS, miRNASNP, MicroSNiPer, and MirSNP, none of them evaluated the influences of tarSNPs on drug response alterations. In spite of the lack of large scale experimentally confirmed tarSNP data, notably, all nine only compared their performance with Mir2Drug using validated data from Table 2. All nine known tarSNPs confirmed in the literature had been predicted in our database.

From twelve cases, only four was corroborated in all four databases. Other four cases were corroborated no more than two databases; in our databases, they all have a certain degree of change in score. One case was not corroborated in all four databases, but it was predicted in our databases. The highest of the free energy change is 10.9 kcal/mol about rs2278414 in ZNF350 3′ UTR and has-miR-21-3p. And rs2278414: C > T was significantly associated with age-related cataract (ARC) risk, which associated with DNA double-strand break repair (DSBR) and nucleotide excision repair (NER) pathway [29].

3.4. Online Database Implementation. We have constructed a user-friendly web tool on genetic variations in miRNA target sites and their potential function in drug response. More specifically, an important focus of this study is to highlight the association of tarSNPs and drug response, thereby identifying tarSNPs that might possibly be involved in drug sensitivity and resistance. We presented a user-friendly website, Mir2Drug database, which will serve as the platform site to provide a practical resource of these drug target-related miRNA polymorphisms and their potential drug response alterations caused by target loss and gain information for all researchers and explore the association of variations in miRNA targets and cancer therapies efficacy and facilitate a mechanistic understanding of relationships among the genetic variations and drugs response. We packaged all the data into a MySQL database and built a user-friendly online website. The Mir2Drug database provides information from the two aspects: (1) drug response alterations medicated by miRNA polymorphisms in target 3′ UTR and (2) drug, pathway, and PPI information about drug targets. Mir2Drug supplies multiple functions for data browsing and searching by search gene symbol, SNP ID, miRNA ID, and an advanced search.

4. Discussion

In this study, we present a database, Mir2Drug, which provides comprehensive annotation information on genetic variations located in miRNA target sites belonging to drug target genes. We evaluate all SNPs in the 3′ UTRs, even if farther away from the miRNA target site, which can alter the miRNA regulation and hence would contribute to drug response. It is appreciated that most of the known cases can be rediscovered in our database. An important goal of this work is to identify the SNPs that can alter miRNA regulations and are also potentially associated with drug sensitivity and resistance in clinical trials. The database would be a valuable resource for experimentalists to explore the functional role of this class of SNP.

Although these SNPs are rare, they may be functionally important, because they can alter extensive mRNA expression by gain or loss of miRNA regulation. Recent studies have reported that genetic variations in miRNA processing genes and miRNA binding sites may affect the biogenesis of

Table 2: Experimental validated miRNA polymorphisms and their predictions in Mir2Drug databases.

SNP id	Gene ID	Gene symbol	miRNA	Allele	Wild free Energy	Mutation free energy	ΔΔ*G* (kcal/mol)	Effect class	miRNASNP	PolymiRTS	MicroSNiPer	MirSNP
rs5186	185	AGTR1	hsa-miR-155-5p	A > C	−8.86	None	−8.86	Complete loss	Yes	Yes	Yes	Yes
rs4245739	4194	MDM4	hsa-miR-191-5p	C > A	−6.07	−0.098	−5.972	Partial loss	Yes	No	Yes	Yes
rs1434536	658	BMPR1B	hsa-miR-125b-5p	C > T	−7.27	−5.08	−2.19	Partial loss	No	No	Yes	Yes
rs3739008	4862	NPAS2	hsa-miR-17-5p	C > T	−9.56	−6.76	−2.8	Partial loss	Yes	Yes	No	Yes
rs3739008	4862	NPAS2	hsa-miR-519e-3p	C > T	−11.71	−8.91	−2.8	Partial loss	Yes	Yes	Yes	Yes
rs3739008	4862	NPAS2	hsa-miR-20b-5p	C > T	−7.49	−4.69	−2.8	Partial loss	Yes	Yes	No	Yes
rs1805672	10219	KLRG1	hsa-miR-584-5p	A > G	−6.9	None	−6.9	Complete loss	No	Yes	Yes	No
rs56109847	285242	HTR3E	hsa-miR-510-5p	G > A	−14.07	−9.69	−4.38	Partial loss	Yes	Yes	Yes	Yes
rs3203358	2626	GATA4	hsa-miR-583	C > G	−10.49	−5.98	−4.51	Partial loss	Yes	No	Yes	No
rs2278414	59348	ZNF350	hsa-miR-21-3p	C > T	None	−10.9	10.9	Complete gain	No	No	Yes	Yes
rs2278414	59348	ZNF350	hsa-miR-150-5p	C > T	−6.22	−10.96	4.74	Partial gain	Yes	Yes	Yes	Yes
rs868	7046	TGFBR1	hsa-let-7b-5p	A > G	−1.93	−0.24	−1.69	Partial loss	No	No	No	No

miRNA and the regulatory effect of miRNAs on their target genes and have a role in cancer development and treatment response [30–34]. However, a few of genes were found on the relationship between SNP in this gene and drug response. We have a new discovery that an SNP (rs751012151) in MRPL4 can be downregulated by hsa-miR-6089 and decrease drug resistance in our database ($\Delta\Delta G$ = 45.54 kcal/mol), but so far it has no literature reported about the relationship between the MRPL4 and drug response. Future studies are necessary to explore the functional role of this class of SNPs. In addition, the known experimentally verified miRNA-disease associations were insufficient in this study. However, in many cases, such information would be very valuable. For example, we can collect this information to quantitatively compare the performance of existing tools (PolymiRTS, MicroSNiPer, MirSNP, and miRNASNP) using cross-validation method if we have sufficient experimentally verified cases and further optimize our method. With the accumulation of such validation and experimental confirmation of miRNA target interactions data, we plan to include this information in next version of our database; we would expect a much better annotation of Mir2Drug in the near future.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

Authors' Contributions

Xianyue Wang, Hong Jiang, and Wei Wu contributed equally to this work.

Acknowledgments

This research was supported by grants from project supported by the Natural Science Foundation of Jiangsu Province, China (Grant nos. BK20131385, BK20161026), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (Grant no. 13KJD520008), and the Natural Science Foundation of China (Grant nos. 81572893, 81502443).

References

[1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of microRNA-target recognition," *PLoS biology.*, vol. 3, no. 3, p. e85, 2005.

[3] N. Meola, V. A. Gennarino, and S. Banfi, "microRNAs and genetic diseases," *PathoGenetics*, vol. 2, no. 1, article 7, 2009.

[4] P. J. Mishra and J. R. Bertino, "MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine," *Pharmacogenomics*, vol. 10, no. 3, pp. 399–416, 2009.

[5] P. Sethupathy and F. S. Collins, "MicroRNA target site polymorphisms and human disease," *Trends in Genetics*, vol. 24, no. 10, pp. 489–497, 2008.

[6] P. J. Mishra, R. Humeniuk, P. J. Mishra, G. S. A. Longo-Sorbello, D. Banerjee, and J. R. Bertino, "A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 33, pp. 13513–13518, 2007.

[7] J. Wynendaele, A. Böhnke, E. Leucci et al., "An illegitimate microRNA target site within the 3′UTR of MDM4 affects ovarian cancer progression and chemosensitivity," *Cancer Research*, vol. 70, no. 23, pp. 9641–9649, 2010.

[8] V. Boni, R. Zarate, J. C. Villa et al., "Role of primary miRNA polymorphic variants in metastatic colon cancer patients treated with 5-fluorouracil and irinotecan," *Pharmacogenomics Journal*, vol. 11, no. 6, pp. 429–436, 2011.

[9] F. Meng, R. Henson, M. Lang et al., "Involvement of human micro-RNA in growth and response to chemotherapy in human cholangiocarcinoma cell lines," *Gastroenterology*, vol. 130, no. 7, pp. 2113–2129, 2006.

[10] L. Bao, M. Zhou, L. Wu et al., "PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits," *Nucleic Acids Research*, vol. 35, no. 1, pp. D51–D54, 2007.

[11] S. Hiard, C. Charlier, W. Coppieters, M. Georges, and D. Baurain, "Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D640–D651, 2009.

[12] M. Barenboim, B. J. Zoltick, Y. Guo, and D. R. Weinberger, "MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets," *Human Mutation*, vol. 31, no. 11, pp. 1223–1232, 2010.

[13] M. Hariharan, V. Scaria, and S. K. Brahmachari, "dbSMR: a novel resource of genome-wide SNPs affecting microRNA mediated regulation," *BMC Bioinformatics*, vol. 10, article 108, 2009.

[14] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D154–D158, 2008.

[15] W. J. Kent, C. W. Sugnet, T. S. Furey et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[16] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nature Genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[17] C. Knox, V. Law, T. Jewison et al., "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1035–D1041, 2011.

[18] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[19] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[20] A. Bhattacharya, J. D. Ziebarth, and Y. Cui, "PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways," *Nucleic Acids Research*, vol. 42, no. 1, pp. D86–D91, 2014.

[21] C. Liu, F. Zhang, T. Li et al., "MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs," *BMC Genomics*, vol. 13, no. 1, article 661, 2012.

[22] J. Gong, Y. Tong, H.-M. Zhang et al., "Genome-wide identification of SNPs in MicroRNA genes and the SNP effects on

MicroRNA target binding and biogenesis," *Human Mutation*, vol. 33, no. 1, pp. 254–263, 2012.

[23] A. Amid, W. D. W. Chik, P. Jamal, and Y. Z. H.-Y. Hashim, "Microarray and quantitative PCR analysis of gene expression profiles in response to treatment with tomato leaf extract in MCF-7 breast cancer cells," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 12, pp. 6319–6325, 2012.

[24] N. Fouz, A. Amid, and Y. Z. H.-Y. Hashim, "Gene expression analysis in MCF-7 breast cancer cells treated with recombinant bromelain," *Applied Biochemistry and Biotechnology*, vol. 173, no. 7, pp. 1618–1639, 2014.

[25] J. V. Vadgama, Z. Scuric, R. Chakrabarti, E. Marzo, D. Shen, and Y. Wu, "Insulin-like growth factor I differentially regulates the expression of HIRF1/hCAF1 and BTG1 genes in human MCF-7 breast cancer cells," *International Journal of Molecular Medicine*, vol. 18, no. 1, pp. 129–139, 2006.

[26] M. Tuna, A. Chavez-Reyes, and A. M. Tari, "HER2/neu increases the expression of Wilms' Tumor 1 (WT1) protein to stimulate S-phase proliferation and inhibit apoptosis in breast cancer cells," *Oncogene*, vol. 24, no. 9, pp. 1648–1652, 2005.

[27] S. S. Ebada, B. Schulz, V. Wray et al., "Arthrinins A-D: novel diterpenoids and further constituents from the sponge derived fungus *Arthrinium* sp.," *Bioorganic and Medicinal Chemistry*, vol. 19, no. 15, pp. 4644–4651, 2011.

[28] C.-H. Chou, N.-W. Chang, S. Shrestha et al., "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Research*, vol. 44, no. 1, pp. D239–D247, 2016.

[29] S. Gu, H. Rong, G. Zhang, L. Kang, M. Yang, and H. Guan, "Functional SNP in 3′-UTR MicroRNA-binding site of ZNF350 confers risk for age-related cataract," *Human Mutation*, vol. 37, no. 11, pp. 1223–1230, 2016.

[30] Y. Ye, K. K. Wang, J. Gu et al., "Genetic variations in MicroRNA-related genes are novel susceptibility loci for esophageal cancer risk," *Cancer Prevention Research*, vol. 1, no. 6, pp. 460–469, 2008.

[31] D. Liang, L. Meyer, D. W. Chang et al., "Genetic variants in microRNA biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response," *Cancer Research*, vol. 70, no. 23, pp. 9765–9776, 2010.

[32] H. Yang, C. P. Dinney, Y. Ye, Y. Zhu, H. B. Grossman, and X. Wu, "Evaluation of genetic variants in microRNA-related genes and risk of bladder cancer," *Cancer Research*, vol. 68, no. 7, pp. 2530–2537, 2008.

[33] G. Sun, J. Yan, K. Noltner et al., "SNPs in human miRNA genes affect biogenesis and function," *RNA*, vol. 15, no. 9, pp. 1640–1651, 2009.

[34] B. M. Ryan, A. I. Robles, and C. C. Harris, "Genetic variation in microRNA networks: the implications for cancer research," *Nature Reviews Cancer*, vol. 10, no. 6, pp. 389–402, 2010.

Hindawi
International Journal of Genomics
Volume 2017, Article ID 1926304, 9 pages
https://doi.org/10.1155/2017/1926304

Review Article

The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies

Pan Zhang,[1] Brian D. Lehmann,[2] Yu Shyr,[3] and Yan Guo[1]

[1]*Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA*
[2]*Department of Biochemistry, Vanderbilt University, Nashville, TN, USA*
[3]*Department of Biostatistics, Vanderbilt University, Nashville, TN, USA*

Correspondence should be addressed to Yan Guo; yan.guo@vanderbilt.edu

Received 28 November 2016; Accepted 9 January 2017; Published 26 January 2017

Academic Editor: Marco Gerdol

High throughput genomic assays empower us to study the entire human genome in short time with reasonable cost. Formalin fixed-paraffin-embedded (FFPE) tissue processing remains the most economical approach for longitudinal tissue specimen storage. Therefore, the ability to apply high throughput genomic applications to FFPE specimens can expand clinical assays and discovery. Many studies have measured the accuracy and repeatability of data generated from FFPE specimens using high throughput genomic assays. Together, these studies demonstrate feasibility and provide crucial guidance for future studies using FFPE specimens. Here, we summarize the findings of these studies and discuss the limitations of high throughput data generated from FFPE specimens across several platforms that include microarray, high throughput sequencing, and NanoString.

1. Introduction

The technique of FFPE is a widely used histological method that uses formalin to fix and paraffin embedding to preserve tissues for extended periods of time. However, the advantages of FFPE processing such as technical ease and low storage cost come at the expense of the sample quality. During the process of fixation, the tissue DNA can be altered by chemical modification, DNA trapping and fragmentation [1, 2], resulting from extensive cross-linking between proteins and nucleic acids [3].

The effects of formalin-fixation are far greater on RNA, as RNA can be altered by severe RNA degradation, chemical modification [4], poly-A tail damage [5], and covalent modification of RNA nucleotide bases by monomethylol (-CH2OH) addition [6]. These covalent modifications can impact reverse transcription from mRNA to cDNA and significantly alter gene expression profiling.

Despite these shortcomings, researchers have successfully been using RNA and DNA extracted from FFPE specimens for high throughput genomic studies. Herein, we review the applications of FFPE specimens in high throughput genomic studies using several technologies, including microarray, high throughput sequencing (HTS), and NanoString.

2. Technical Challenges and Concerns

One major challenge in using FFPE specimens in genomic assays is the low quality and quantity of nucleic acids extracted from FFPE blocks. The process of FFPE is designed to well preserve cellular proteins that can be evaluated by immunohistochemistry based assays rather than preserving DNA or RNA. It is known that formalin-fixation can cause nucleic acids fragmentation, degradation, and cross-linking to proteins [1–3, 7–11]. In addition, the long storage time can further compromise the quality of the nucleic acids within FFPE blocks [12]. Nucleic acid degradation and crosslinking to proteins can significantly reduce the quantity of DNA and RNA extracted from FFPE specimens, while nucleic acids fragmentation can reduce library fragment size and uniformity. Further complicating genomic assay is the limited quality control measurements that are performed on FFPE specimens such as traditional RNA integrity number (RIN) measurements that do not truly reflect the success chance

of sequencing from FFPE specimens [13]. Together, reduced quantity and integrity of the extracted nucleic acids can lead to the failure of HTS library construction.

Prior to sequencing, commercially available kit, such as Illumina TruSeq, is required to assemble a sequencing library. Most commercial kits have similar performance. Unlike DNA, there are several methods to enrich for RNA prior to library construction that use depletion of highly abundant ribosomal RNA or oligo-dT to capture mRNAs with polyA tails. For RNA library construction, ribosomal RNA depletion method is preferred to oligo dT capture because many mRNA transcripts from FFPE specimens have lost their polyA tails from to extensive fragmentation [14]. Both Illumina and SOLiD HTS platforms have been demonstrated to work well with FFPE derived libraries and different platforms does not seem to have a bias toward FFPE specimens [15, 16]. While HTS libraries have been constructed from nucleic acids with poor quality, those studies [17, 18] have shown that the sequencing data generated were less than ideal quality.

3. Microarray

Gene expression microarray uses large-scale arrays of fluorescent oligonucleotide probes to measure mRNA expression across many genes simultaneously and was the driving force for high throughput gene expression studies prior to the introduction of RNA-seq. During the gene expression microarray era, FFPE specimens had been extensively used for expression profiling purposes [19–22]. Because the quality of RNA extract from FFPE specimens is always of questionable quality, many studies [23–28] were conducted to evaluate the integrity of FFPE gene expression microarray data by comparing the gene expression consistency between paired FFPE and fresh frozen (FF) samples. All of the comparative studies have found that reasonable consistency of gene expression quantified from FFPE and FF specimens likely attributed the oligonucleotide probes measure expression being located at several positions across a gene. In addition to mRNA transcript quantification, microarray technology has been adapted to measure DNA copy number, single nucleotide polymorphisms (SNPs), and DNA methylation.

The most frequent types of variation in the genome are single base differences between two DNA sequences and genotyping microarray has been developed to detect single nucleotide polymorphisms in genomic DNA. Although DNA is more stable than RNA, the quality of DNA extracted from FFPE specimens can be considerably compromised by artefactual nucleotide changes introduced by formalin-fixation. Therefore, many studies have evaluated the feasibility of using DNA extracted from FFPE specimens for genotyping array analysis [29–33]. These studies have shown a high concordance in SNP calls between FF and FFPE specimens. Encouraged by these findings, researchers have widely used FFPE specimens in a variety of genotyping array studies [30, 31, 34–37]. In addition to SNP detection, genotyping arrays can also be used to estimate DNA copy number variance (CNV). However, CNV estimation from DNA obtained from FFPE specimens can be challenging, as DNA usually degraded and fragmented. Nonetheless, several modified protocols have been reported and different array platforms have been tested for the practicability of performing CNV analysis with FFPE specimens [29, 33, 36]. All these studies show plausible concordance of CNVs identified between paired FFPE and FF specimens.

In addition to CNV estimation from genotype arrays, comparative genomic hybridization (CGH) arrays have been developed as a genome-wide, high-resolution technique for the detection of copy number variations between two genomes. As aforementioned, CNV detection is more susceptible to the fragmented nature of DNA extracted from FFPE specimens. One study has shown that FFPE specimens can have spurious copy number variation in array-CGH profiles [38]. For successful CNV estimation from array-CGH, several requirements for DNA have been suggested for FFPE [39]. First, it was found that only FFPE tissues that supported polymerase chain reaction (PCR) amplification of >300 bp DNA fragment provided high quality, reproducible array-CGH data. Second, roughly 10 ng DNA from FFPE tissues is needed as input for array-CGH analysis prior to whole genome amplification. Third, high tumor cellularity of greater than 70% tumor DNA was required for reliable array-CGH analysis [39].

Prior to hybridization, DNA must undergo whole genome DNA amplification and several amplification methods can also affect the quality of array-CGH data [40]. Random-primed amplification was found to be superior to degenerate oligonucleotide-primed amplification [40]. Several studies have proposed optimized protocols for array-CGH analysis using DNA from FFPE specimens [41, 42]. Comparison studies using either paired FF specimens or fluorescent in situ hybridization (FISH) methods as a gold standard have demonstrated that array-CGH are reliable for CNV estimation from FFPE specimens [43–45]. This reliability has allowed for a clinical application of array-CGH to distinguish Spitz nevus and melanoma in FFPE specimens [46].

DNA can be modified by several mechanisms that can alter gene transcription including methylation of CpG sites and microarray technologies have been adapted to measure global methylation patterns of DNA. These methods largely rely on bisulfite treatment to convert unmethylated cytosine to uracil and the latest methylation EPIC BeadChips from Illumina can interrogate over 850,000 CpG sites at single nucleotide resolution. Several studies compared methylation values measured from Illumina methylation arrays on paired FFPE and FF specimens and found high level of concordance ($R^2 > 0.95$) [47–50]. While study did report lower concordance between FFPE and FF specimens ($r = 0.6$) [51], others have questioned the statistical considerations and batch effect that may have impacted this study [52]. The overall good performance of FFPE in methylation arrays is likely due to the better stability of DNA compared to RNA. To date, many epigenetic methylation studies have used FFPE specimens as their source [53, 54].

4. RNA-Seq

With the rise of HTS technology, RNA-seq has inevitably replaced microarray as the platform of choice for expression

profiling technology [55–59]. RNA-seq provides numerous advantages over microarray technology, including the identification of all RNAs in the library rather than RNA with predesigned probes, allowing the expression quantification at multiple levels (gene, transcript, and exon) without designing specific probes and permitting the additional discovery opportunities such as gene fusion and allelic specific expression.

Similar to microarray technology, FF tissue samples provide the highest data quality. However, majority of specimens are processed by FFPE and researchers have been applying the same strategy as during the microarray era, evaluating the accuracy and repeatability of gene quantification using HTS technology by comparing matched pairs of FF and FFPE specimens from the same subject.

Norton et al. calculated the correlation of gene expression across nine matched pairs of FF and FFPE specimens and Pearson correlations ranged from 0.60 to 0.83 [60]. Graw et al. analyzed RNA-seq data from six pairs of FF/FFPE tumor samples and found that the correlations of gene expression data were greater than 0.89. The same study also reported 99.67% concordance between sequence variations identified from FFPE RNA and FF DNA [61]. In Hester et al.'s study, storage time was shown to impact concordance between paired FF and FFPE specimens with high concordance of specimens stored less than 2 years ($r^2 = 0.99$) compared to FFPE specimens with storage greater than 20 years ($r^2 = 0.84$) [62]. Hedegaard et al. compared the expression profiles from 27 FFPE and FF pairs from different tissues (colon, bladder, and prostate) and with different storage time. The results revealed a high degree of Pearson correlation ($r > 0.90$) across all pairs [18]. Zhao et al. used two ribosomal RNA removal kits (Ribo Zero and Duplex-Specific Nuclease) to sequence paired FFPE and FF specimens. Both protocols resulted in a Pearson correlation of about 0.90 between matched pair of FFPE and FF specimens [63]. Eikrem et al. compared the gene expression profiles across 16 pairs of FF and FFPE specimens and the correlation of the average expression is 0.97 [64]. Li et al. also reported a correlation more than 0.91 between FF and FFPE pairs [65]. These studies show that reliable gene expression data can be obtained from whole transcriptome sequencing of FFPE specimens; provided tissues blocks have not been stored from long periods.

In addition to gene expression quantification, RNA-seq data can be mined for single nucleotide variants and structural alterations such as gene rearrangements that result in hybrid transcripts [66]. However, unlike gene expression quantification, these additional data mining opportunities do not apply well for RNA-seq data generated from FFPE specimens. One comparative study found that only 24% of high-confidence fusion transcripts detected in FF specimens were also detected in matched FFPE specimens [60]. This low recovery rate occurs despite threefold increases sequencing depth. Another study found that between SNVs identified from RNA-seq replicates from FFPE specimens showed extremely poor genotype consistency (<50%), rendering it unreliable for SNV detection [14].

Thus far, overwhelming findings provided emerging evidence of the accurate expression profiles obtained from FFPE specimens; an increasing number of studies began to use RNA-seq technology on FFPE specimens to perform gene expression profiling [67–75]. While gene expression quantification has produced reliable results, other data mining opportunities such as gene fusion and SNV detection have been found to be not feasible with FFPE specimens.

5. Small RNA-Seq

MicroRNAs (miRNA) are small noncoding RNA molecules containing around 22 nucleotides and have been found to play an important role in many biological processes. MiRNAs function through base-pairing with complementary sequences within mRNA molecules and these mRNA molecules are subsequently silenced. HTS has also revolutionized the miRNA research area. Compared to traditional methods such as TaqMan gene expression assay and microarray, HTS enables the detection of almost all small RNAs present in the samples, including novel and underexpressed miRNAs as well as small RNAs of other categories [76].

Since miRNAs are more stable than RNA molecules [77–79], HTS is quite promising for quantifying miRNA profiles from FFPE specimens. Several pioneering studies using matched FF and FFPE specimens have already been performed to evaluate the usefulness of FFPE specimen for miRNA-seq technology. These studies have found that miRNA-seq data generated from FFPE specimens have similar number of total reads but tend to have a slightly shorter average read length after trimming for adapter sequences [80–83].

In addition, the proportion of reads that can be mapped to miRNAs was also lower in FFPE specimens [80, 81]. The decreased mapping could be due to small fragments of other RNA species such as degraded lncRNAs and mRNAs in the small RNA library [81]. Most studies agree that the small RNAs composition from FFPE specimens is similar to that from FF specimens [81, 83], and correlations between miRNA expression levels quantified from paired FF and FFPE specimens range from 0.71 to 0.98 [80, 81, 83]. More interestingly, against common intuition, two studies found that storage time of the FFPE blocks did not affect the quality of miRNA-seq data [81, 83]. These studies further showed that while the total miRNA expression profile is highly correlated between matched FF and FFPE specimens, the relative read count of each miRNA is dependent on GC content. Specifically, GC-poor miRNAs were shown to be more degraded than GC-rich miRNAs [80].

Encouraged by these validation studies, researchers began to apply HTS miRNA-seq to FFPE specimens [84, 85]. Plieskatt et al. applied miRNA-seq on FFPE preserved nasopharyngeal carcinoma tissues. They found that FFPE tissue can yield RNA of sufficient quality for downstream sequencing analysis. Using the miRNA profile generated from these FFPE specimens, the authors identified Epstein Barr Virus miRNAs as potential NPC biomarkers [84]. Riester et al. collected 16 osteosarcomas FFPE specimens and 14 osteoblastomas FFPE specimens. miRNA-seq analysis of

these 30 FFPE specimens allows the authors to identify miR-210 as a discriminatory marker that distinguishes between osteoblastoma and osteosarcoma [85].

6. DNA-Seq

HTS technologies have been widely used to characterize variations and quantity of DNA from both normal and diseased tissue. DNA-sequencing can be used to characterize genomic variants such as SNV, insertions/deletions (Indels), copy number variations (CNVs), and structural gene rearrangements. HTS DNA-seq performs better with high quality DNA from FF specimens as starting materials. However, FFPE specimens have also been evaluated using DNA-seq.

Similar to comparisons of microarray and RNA-seq, many studies have used matched paired FFPE and FF specimens to evaluate the quality of genomic variants identified from FFPE specimens. The overall concordance of SNV calls between FF and FFPE specimens across different studies ranges from 70% to 99.8% [15–18, 86–92]. In most cases, more than 80% of SNVs identified in FF specimens can be reliably recovered from the matched FFPE specimens. Furthermore, many studies found that a significantly higher number of unique SNVs can be identified from FFPE specimens than matched FF specimens and likely attributed to chemical modification of nucleotides by formalin-fixation. Specifically, formalin-fixation can cause deamination of cytosine bases to uracil. Thus, during amplification, if DNA polymerase reads across a uracil change, artefactual C>T/G>A changes can occur and introduce false positives [10]. Kerick et al. found that among the 149 false positives SNV calls from a FFPE specimen, all but four can be explained by the fixation process [88]. As an alternative, uracil-DNA glycosylase (UDG) was reported to be used to remove uracil-containing deaminated DNA molecules before library construction and treatment reduces C>T and A>G variant calls by 77% and 94%, respectively [93]. While FFPE specimens have a higher rate of nonreproducible SNVs, their random distributions allow for increased coverage to diminish the false positive rate [89]. One study showed that increasing sequencing coverage to 80x reduced significantly the false positive rate and increased the concordance between FF and FFPE specimens [88]. However, the depth of sequencing to produce reliable SNV calls is unrealistic for most whole genome sequencing and whole exome sequencing analysis.

Similar to SNV detection, FFPE specimens have also been evaluated for their feasibility for insertions and deletions (indel) detection. The concordance of indel calls between FFPE specimens and matched FF specimens has been mixed, ranging from 62% to 98.25% [88, 89, 91]. CNV estimations have also been inconsistent among studies with DNA-seq from FFPE specimens. Using whole genome sequencing, Schweiger et al. reported that the CNVs found were identical for FF and FFPE specimens [16]. However, Menon et al. used whole exome sequencing and reported that there is a high degree of noise in CNV calling from FFPE specimens, probably due to DNA degradation [15]. Munchel et al. used low-pass whole genome sequencing and found that the CNVs within segmented regions between paired FF and FFPE specimens are similar although the size of predicted CNVs differed between paired samples [89]. Several factors may have contributed to the relatively poor concordance of CNV calls between FF and FFPE specimens. First, FFPE specimens tend to have a high degree of cellular heterogeneity. A low purity of tumor cells or the presence of substantial immune cells can make CNV estimations noisy from FFPE specimens. Isolating pure population of tumor cells from FFPE specimen by flow cytometry based methods may circumvent this issue and improve CNV detection [87]. Another potential explanation for high CNV variation may stem from comparisons using lower coverage [89].

Together, these studies provide convincing evidence that accurate SNV can be identified from DNA-seq data from FFPE specimens and many studies have already taken advantage of large FFPE repository with DNA-seq technology to drive new scientific discoveries [93–103].

7. Applications in Other Type of HTS

DNA-seq has been modified to measure global DNA methylation patterns similar to methylation arrays using bisulfite treatment of DNA. Although less popular than DNA and RNA-seq, there have been successful usages of FFPE in bisulfite sequencing [104, 105]. One study evaluated the practicability of using FFPE specimens in bisulfite sequencing and found that the correlation between paired FFPE and FF specimens was good ($r = 0.87$) [106]. Several protocols and methodologies for bisulfite sequencing of FFPE specimens have been established [107, 108].

Chromatin immunoprecipitation sequencing (ChIP-seq) is a form of HTS that can identify global binding sites of DNA associated proteins. The usage of FFPE specimens for ChIP-seq can be difficult due to limited isolation of soluble DNA-protein complexes that are altered by excessive chemical cross-linking during formalin-fixation process. However, Fanelli et al. published a protocol, which demonstrated successful identification of DNA-protein binding sites using FFPE specimens [109]. This protocol has yet to be adapted widely for the usage of FFPE specimens. In 2016, Cejas et al. proposed a fixed-tissue chromatin immunoprecipitation sequencing (FiT-seq), which enables reliable extraction of soluble chromatin from FFPE specimens [110]. Whether this method will be more received by the research community remains to be seen. There are other types of HTS such as nuclear run-on assay (GRO-seq or PRO-seq) and cross-linking immunoprecipitation sequencing (CLIP-seq). These types of applications of HTS have not been used to the extent of DNA- and RNA-seq; thus few studies have been done using FFPE specimens.

8. NanoString

Similar to microarray technology, the NanoString nCounter system can directly measure gene expression by using multiplexed color-coded probe-pairs and offers high levels of precision and sensitivity (<1 copy per cell). The technology uses molecular "barcodes" and single molecule imaging to detect and count hundreds of unique transcripts in a single reaction.

Because nCounter system is quantitative and does not require reverse transcription and amplification, it is free from any bias and errors introduced by the reverse transcription and the amplification processes. This is also the major reason for the claim that NanoString nCounter technology works well with FFPE specimens [111]. Naturally, several studies also investigated the performance of NanoString on FFPE specimens.

An original study conducted by NanoString company from 2008 measured concordance of gene expression measured by NanoString and RT-PCR/microarray and found high correlations (RT-PCR $R^2 = 0.79$, Microarray $R^2 = 0.95$). However, several additional follow-up studies found only moderate correlation between NanoString and RT-PCR, with correlation ranging from 0.48 to 0.59 [112–114]. This level of correlation holds true for both mRNA and miRNA measurement. In addition, the concordance of NanoString with other high throughput platforms, such as microarray and HTS, was also less than ideal, with correlations around 0.5 [14, 115–117]. On a positive note, NanoString was used with FFPE specimens to subtype diffuse large B-cell lymphoma [118]. The subtyping results by nCounter system have a 90% concordance rate with the results generated by Hans immunohistochemistry [118]. Based on the overall evidence presented thus far, we are not yet convinced that NanoString nCounter system is the definite technology for measuring gene expression from FFPE specimens. One of the major limitations of NanoString is that it is not a true high throughput technology, measuring up to a few hundred genes that have been chosen with prior knowledge. However, the limited throughput of NanoString is efficient enough to perform clinical assays such as Prosigna Panel and MammaPrint.

9. Discussion

FFPE processing of tissue is not the most ideal method for quantifying RNA and DNA variations with HTS methods. However, it is often chosen over FF storage because of minimal cost and ease of storage. With high throughput genomic assays dominating the biomedical research field, the ability to expand these studies to existing large FFPE specimen repositories can accelerate and rapidly verify discoveries. Numerous studies have been conducted to evaluate the performance of FFPE specimens with high throughput assays, including gene expression microarray, genotyping microarray, aCGH, methylation array, RNA-seq, DNA-seq, bisulfite sequencing, ChIP-seq, and NanoString. Together the current studies have established that FFPE can generate reliable data for gene expression and SNV detection. However, for more complex alterations such as indel, CNV estimation, and detection of hybrid transcripts, FFPE specimens have been proven to be less than ideal. The overall consensus for utilizing FFPE specimens in high throughput genomic study is that the data quality is negatively correlated to storage time. However, small RNAs have been shown to be an exception to this rule, due to the already small size of the small RNA which is less affected by the degradation of RNA.

Overall, FFPE specimens provide great value in biomedical research and can be utilized for HTS applications. However, there is always a high risk associated FFPE specimen based high throughput genomic assays because the quality of the FFPE specimens is near impossible to determine. Thus, a small pilot studies should be considered to establish feasibility prior to committing resources to a large FFPE based study.

Competing Interests

The authors declare that they have no competing interests.

References

[1] M. T. P. Gilbert, T. Haselkorn, M. Bunce et al., "The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when?" *PLoS ONE*, vol. 2, no. 6, article no. e537, 2007.

[2] I. Daugaard, T. E. Kjeldsen, H. Hager, L. L. Hansen, and T. K. Wojdacz, "The influence of DNA degradation in formalin-fixed, paraffin-embedded (FFPE) tissue on locus-specific methylation assessment by MS-HRM," *Experimental and Molecular Pathology*, vol. 99, no. 3, pp. 632–640, 2015.

[3] F. Lewis, N. J. Maughan, V. Smith, K. Hillan, and P. Quirke E, "Unlocking the archive—gene expression in paraffin-embedded tissue," *Journal of Pathology*, vol. 195, no. 1, pp. 66–71, 2001.

[4] J.-Y. Chung, T. Braunschweig, and S. M. Hewitt, "Optimization of recovery of RNA from formalin-fixed, paraffin-embedded tissue," *Diagnostic Molecular Pathology*, vol. 15, no. 4, pp. 229–236, 2006.

[5] M. D. McKinney, S. J. Moon, D. A. Kulesh, T. Larsen, and R. J. Schoepp, "Detection of viral RNA from paraffin-embedded tissues after prolonged formalin fixation," *Journal of Clinical Virology*, vol. 44, no. 1, pp. 39–42, 2009.

[6] N. Masuda, T. Ohnishi, S. Kawamoto, M. Monden, and K. Okubo, "Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples," *Nucleic Acids Research*, vol. 27, no. 22, pp. 4436–4443, 1999.

[7] C. Auerbach, M. Moutschen-Dahmen, and J. Moutschen, "Genetic and cytogenetical effects of formaldehyde and related compounds," *Mutation Research/Reviews in Genetic Toxicology*, vol. 39, no. 3-4, pp. 317–361, 1977.

[8] M. Y. Feldman, "Reactions of nucleic acids and nucleodroteins with formaldehyde," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 13, pp. 1–49, 1973.

[9] F. Karlsen, M. Kalantari, M. Chitemerere, B. Johansson, and B. Hagmar, "Modifications of human and viral deoxyribonucleic acid by formaldehyde fixation," *Laboratory Investigation*, vol. 71, no. 4, pp. 604–611, 1994.

[10] M. Srinivasan, D. Sedmak, and S. Jewell, "Effect of fixatives and tissue processing on the content and integrity of nucleic acids," *The American Journal of Pathology*, vol. 161, no. 6, pp. 1961–1971, 2002.

[11] S. von Ahlfen, A. Missel, K. Bendrat, and M. Schlumpberger, "Determinants of RNA quality from FFPE samples," *PLoS ONE*, vol. 2, no. 12, Article ID e1261, 2007.

[12] D. M. Carrick, M. G. Mehaffey, M. C. Sachs et al., "Robustness of next generation sequencing on older formalin-fixed paraffin-embedded tissue," *PLoS ONE*, vol. 10, no. 7, Article ID e0127353, 2015.

[13] Q. Sheng, K. Vickers, S. Zhao et al., "Multi-perspective quality control of Illumina RNA sequencing data analysis," *Briefings in Functional Genomics*, 2016.

[14] Y. Guo, J. Wu, S. Zhao et al., "RNA sequencing of formalin-fixed, paraffin-embedded specimens for gene expression quantification and data mining," *International Journal of Genomics*, vol. 2016, Article ID 9837310, 10 pages, 2016.

[15] R. Menon, M. Deng, D. Boehm et al., "Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue," *International Journal of Molecular Sciences*, vol. 13, no. 7, pp. 8933–8942, 2012.

[16] M. R. Schweiger, M. Kerick, B. Timmermann et al., "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number-and mutation-analysis," *PLoS ONE*, vol. 4, no. 5, Article ID e5548, 2009.

[17] A. Astolfi, M. Urbini, V. Indio et al., "Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST)," *BMC Genomics*, vol. 16, no. 1, article no. 892, 2015.

[18] J. Hedegaard, K. Thorsen, M. K. Lund et al., "Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue," *PLoS ONE*, vol. 9, no. 5, Article ID e98187, 2014.

[19] S. Lassmann, C. Kreutz, A. Schoepflin, U. Hopt, J. Timmer, and M. Werner, "A novel approach for reliable microarray analysis of microdissected tumor cells from formalin-fixed and paraffin-embedded colorectal cancer resection specimens," *Journal of Molecular Medicine*, vol. 87, no. 2, pp. 211–224, 2009.

[20] D. Abdueva, M. Wing, B. Schaub, T. Triche, and E. Davicioni, "Quantitative expression profiling in formalin-fixed paraffin-embedded samples by Affymetrix microarrays," *Journal of Molecular Diagnostics*, vol. 12, no. 4, pp. 409–417, 2010.

[21] K. M. Linton, Y. Hey, E. Saunders et al., "Acquisition of biologically relevant gene expression data by affymetrix microarray analysis of archival formalin-fixed paraffin-embedded tumours (vol 98, pg 1403, 2008)," *British Journal of Cancer*, vol. 99, no. 2, article 383, 2008.

[22] M. A. Walter, D. Seboek, P. Demougin et al., "Extraction of high-integrity RNA suitable for microarray gene expression analysis from long-term stored human thyroid tissues," *Pathology*, vol. 38, no. 3, pp. 249–253, 2006.

[23] G. Fedorowicz, S. Guerrero, T. D. Wu, and Z. Modrusan, "Microarray analysis of RNA extracted from formalin-fixed, paraffin-embedded and matched fresh-frozen ovarian adenocarcinomas," *BMC Medical Genomics*, vol. 2, article 23, 2009.

[24] M. Frank, C. Döring, D. Metzler, S. Eckerle, and M.-L. Hansmann, "Global gene expression profiling of formalin-fixed paraffin-embedded tumor samples: a comparison to snap-frozen material using oligonucleotide microarrays," *Virchows Archiv*, vol. 450, no. 6, pp. 699–711, 2007.

[25] M. S. Scicchitano, D. A. Dalmas, M. A. Bertiaux et al., "Preliminary comparison of quantity, quality, and microarray performance of RNA extracted from formalin-fixed, paraffin-embedded, and unfixed frozen tissue samples," *Journal of Histochemistry & Cytochemistry*, vol. 54, no. 11, pp. 1229–1237, 2006.

[26] W. H. Bradley, K. Eng, M. Le, A. C. Mackinnon, C. Kendziorski, and J. S. Rader, "Comparing gene expression data from formalin-fixed, paraffin embedded tissues and qPCR with that from snap-frozen tissue and microarrays for modeling outcomes of patients with ovarian carcinoma," *BMC Clinical Pathology*, vol. 15, article 17, 2015.

[27] L. Mittempergher, J. J. de Ronde, M. Nieuwland et al., "Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue," *PLoS ONE*, vol. 6, no. 2, Article ID e17163, 2011.

[28] L. Roberts, J. Bowers, K. Sensinger, A. Lisowski, R. Getts, and M. G. Anderson, "Identification of methods for use of formalin-fixed, paraffin-embedded tissue samples in RNA expression profiling," *Genomics*, vol. 94, no. 5, pp. 341–348, 2009.

[29] S. Jacobs, E. R. Thompson, Y. Nannya et al., "Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays," *Cancer Research*, vol. 67, no. 6, pp. 2544–2551, 2007.

[30] E. H. Lips, J. W. F. Dierssen, R. Van Eijk et al., "Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays," *Cancer Research*, vol. 65, no. 22, pp. 10188–10191, 2005.

[31] E. R. Thompson, S. C. Herbert, S. M. Forrest, and I. G. Campbell, "Whole genome SNP arrays using DNA derived from formalin-fixed, paraffin-embedded ovarian tumor tissue," *Human Mutation*, vol. 26, no. 4, pp. 384–389, 2005.

[32] H. I. Vos, T. Van Der Straaten, M. J. H. Coenen, U. Flucke, D. M. W. M. Te Loo, and H.-J. Guchelaar, "High-quality genotyping data from formalin-fixed, paraffin-embedded tissue on the drug metabolizing enzymes and transporters plus array," *The Journal of Molecular Diagnostics*, vol. 17, no. 1, pp. 4–9, 2015.

[33] Y. Wang, V. E. Carlton, G. Karlin-Neumann et al., "High quality copy number and genotype data from FFPE samples using Molecular Inversion Probe (MIP) microarrays," *BMC Medical Genomics*, vol. 2, no. 1, article 8, 2009.

[34] Y. Wang, V. E. Carlton, G. Karlin-Neumann et al., "High quality copy number and genotype data from FFPE samples using Molecular Inversion Probe (MIP) microarrays," *BMC Medical Genomics*, vol. 2, article no. 8, 2009.

[35] M. Tuefferd, A. De Bondt, I. Van Den Wyngaert et al., "Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays," *Genes Chromosomes and Cancer*, vol. 47, no. 11, pp. 957–964, 2008.

[36] J. Oosting, E. H. Lips, R. Van Eijk et al., "High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays," *Genome Research*, vol. 17, no. 3, pp. 368–376, 2007.

[37] K. Alvarez, S. F. Kash, M. A. Lyons-Weiler et al., "Reproducibility and performance of virtual karyotyping with SNP microarrays for the detection of chromosomal imbalances in formalin-fixed paraffin-embedded tissues," *Diagnostic Molecular Pathology*, vol. 19, no. 3, pp. 127–134, 2010.

[38] E. A. Mc Sherry, A. Mc Goldrick, E. W. Kay, A. M. Hopkins, W. Gallagher, and P. A. Dervan, "Formalin-fixed paraffin-embedded clinical tissues show spurious copy number changes in array-CGH profiles," *Clinical Genetics*, vol. 72, no. 5, pp. 441–447, 2007.

[39] N. A. Johnson, R. A. Hamoudi, K. Ichimura et al., "Application of array CGH on archival formalin-fixed paraffin-embedded tissues including small numbers of microdissected cells," *Laboratory Investigation*, vol. 86, no. 9, pp. 968–978, 2006.

[40] S. DeVries, S. Nyante, J. Korkola et al., "Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors," *The Journal of Molecular Diagnostics*, vol. 7, no. 1, pp. 65–71, 2005.

[41] S. A. Joosse, E. H. van Beers, and P. M. Nederlof, "Automated array-CGH optimized for archival formalin-fixed, paraffin-embedded tumor material," *BMC Cancer*, vol. 7, article no. 43, 2007.

[42] S. E. Little, R. Vuononvirta, J. S. Reis-Filho et al., "Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA," *Genomics*, vol. 87, no. 2, pp. 298–306, 2006.

[43] H. Fensterer, B. Radlwimmer, J. Sträter et al., "Matrix-comparative genomic hybridization from multicenter formalin-fixed paraffin-embedded colorectal cancer tissue blocks," *BMC Cancer*, vol. 7, article no. 58, 2007.

[44] P. A. Lennon, Y. Zhuang, D. Pierson et al., "Bacterial artificial chromosome array-based comparative genomic hybridization using paired formalin-fixed, paraffin-embedded and fresh frozen tissue specimens in multiple myeloma," *Cancer*, vol. 115, no. 2, pp. 345–354, 2009.

[45] G. Mohapatra, R. A. Betensky, E. R. Miller et al., "Glioma test array for use with formalin-fixed, paraffin-embedded tissue: array comparative genomic hybridization correlates with loss of heterozygosity and fluorescence in situ hybridization," *The Journal of Molecular Diagnostics*, vol. 8, no. 2, pp. 268–276, 2006.

[46] J. D. Harvell, S. Kohler, S. Zhu, T. Hernandez-Boussard, J. R. Pollack, and M. Van De Rijn, "High-resolution array-based comparative genomic hybridization for distinguishing paraffin-embedded Spitz nevi and melanomas," *Diagnostic Molecular Pathology*, vol. 13, no. 1, pp. 22–25, 2004.

[47] S. Moran, M. Vizoso, A. Martinez-Cardús et al., "Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the Infinium HumanMethylation450 Microarray," *Epigenetics*, vol. 9, no. 6, pp. 829–833, 2014.

[48] C. Thirlwell, M. Eymard, A. Feber et al., "Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip," *Methods*, vol. 52, no. 3, pp. 248–254, 2010.

[49] J. K. Killian, S. Bilke, S. Davis et al., "Large-scale profiling of archival lymph nodes reveals pervasive remodeling of the follicular lymphoma methylome," *Cancer Research*, vol. 69, no. 3, pp. 758–764, 2009.

[50] T. C. de Ruijter, J. P. de Hoon, J. Slaats et al., "Formalin-fixed, paraffin-embedded (FFPE) tissue epigenomics using Infinium HumanMethylation450 BeadChip assays," *Laboratory Investigation*, vol. 95, no. 7, pp. 833–842, 2015.

[51] F. Jasmine, R. Rahaman, S. Roy et al., "Interpretation of genome-wide infinium methylation data from ligated DNA in formalin-fixed, paraffin-embedded paired tumor and normal tissue," *BMC Research Notes*, vol. 5, article 117, 2012.

[52] C. Thirlwell, A. Feber, M. Lechner, A. E. Teschendorff, and S. Beck, "Comments on: interpretation of genome-wide infinium methylation data from ligated DNA in formalin-fixed paraffin-embedded paired tumor and normal tissue," *BMC Research Notes*, vol. 5, article no. 631, 2012.

[53] M. G. Kibriya, M. Raza, F. Jasmine et al., "A genome-wide DNA methylation study in colorectal carcinoma," *BMC Medical Genomics*, vol. 4, article no. 50, 2011.

[54] T. D. Dumenil, L. F. Wockner, M. Bettington et al., "Genome-wide DNA methylation analysis of formalin-fixed paraffin embedded colorectal cancer tissue," *Genes Chromosomes & Cancer*, vol. 53, no. 7, pp. 537–548, 2014.

[55] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[56] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[57] Y. W. Asmann, E. W. Klee, E. A. Thompson et al., "3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer," *BMC genomics*, vol. 10, p. 531, 2009.

[58] N. Cloonan, A. R. R. Forrest, G. Kolle et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[59] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.

[60] N. Norton, Z. Sun, Y. W. Asmann et al., "Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors," *PLoS ONE*, vol. 8, no. 11, Article ID e81925, 2013.

[61] S. Graw, R. Meier, K. Minn et al., "Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples," *Scientific Reports*, vol. 5, Article ID 12335, 2015.

[62] S. D. Hester, V. Bhat, B. N. Chorley et al., "Editor's highlight: dose–response analysis of RNA-Seq profiles in archival formalin-fixed paraffin-embedded samples," *Toxicological Sciences*, vol. 154, no. 2, pp. 202–213, 2016.

[63] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, "Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling," *BMC Genomics*, vol. 15, no. 1, article 419, 2014.

[64] O. Eikrem, C. Beisland, K. Hjelle et al., "Transcriptome sequencing (RNAseq) enables utilization of formalin-fixed, paraffin-embedded biopsies with clear cell renal cell carcinoma for exploration of disease biology and biomarker development," *PLoS ONE*, vol. 11, no. 2, Article ID e0149743, 2016.

[65] P. Li, A. Conley, H. Zhang, and H. L. Kim, "Whole-Transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq," *BMC Genomics*, vol. 15, no. 1, article no. 1087, 2014.

[66] L. Han, K. C. Vickers, D. C. Samuels, and Y. Guo, "Alternative applications for distinct RNA sequencing strategies," *Briefings in Bioinformatics*, vol. 16, no. 4, pp. 629–639, 2014.

[67] K. J. Brayer, C. A. Frerich, H. Kang, and S. A. Ness, "Recurrent fusions in MYB and MYBL1 define a common, transcription factor-driven oncogenic pathway in salivary gland adenoid cystic carcinoma," *Cancer Discovery*, vol. 6, no. 2, pp. 176–187, 2016.

[68] L. Erdem-Eraslan, M. J. Van Den Bent, Y. Hoogstrate et al., "Identification of patients with recurrent glioblastoma who may benefit from combined bevacizumab and CCNU Therapy: a report from the BELOB Trial," *Cancer Research*, vol. 76, no. 3, pp. 525–534, 2016.

[69] P.-A. Just, F. Letourneur, C. Pouliquen et al., "Identification by FFPE RNA-Seq of a new recurrent inversion leading to RBM10-TFE3 fusion in renal cell carcinoma with subtle TFE3 break-apart FISH pattern," *Genes Chromosomes and Cancer*, vol. 55, no. 6, pp. 541–548, 2016.

[70] X. S. Lin, L. Hu, K. Sandy et al., "Differentiating progressive from nonprogressive T1 bladder cancer by gene expression profiling: applying RNA-sequencing analysis on archived specimens," *Urologic Oncology*, vol. 23, no. 3, pp. 327–336, 2014.

[71] Y. Liu, A. P. Noon, E. Aguiar Cabeza et al., "Next-generation RNA sequencing of archival formalin-fixed paraffin-embedded urothelial bladder cancer," *European Urology*, vol. 66, no. 6, pp. 982–986, 2014.

[72] Y. Ma, R. Ambannavar, J. Stephans et al., "Fusion transcript discovery in formalin-fixed paraffin-embedded human breast cancer tissues reveals a link to tumor progression," *PLoS ONE*, vol. 9, no. 4, Article ID 0094202, 2014.

[73] M. L. Morton, X. Bai, C. R. Merry et al., "Identification of mRNAs and lincRNAs associated with lung cancer progression using next-generation RNA sequencing from laser microdissected archival FFPE tissue specimens," *Lung Cancer*, vol. 85, no. 1, pp. 31–39, 2014.

[74] D. Sinicropi, K. Qu, F. Collin et al., "Whole transcriptome RNA-seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue," *PLoS ONE*, vol. 7, no. 7, Article ID e40092, 2012.

[75] C. Walther, J. Hofvander, J. Nilsson et al., "Gene fusion detection in formalin-fixed paraffin-embedded benign fibrous histiocytomas using fluorescence in situ hybridization and RNA sequencing," *Laboratory Investigation*, vol. 95, no. 9, pp. 1071–1076, 2015.

[76] K. C. Vickers, L. A. Roteta, H. Hucheson-Dilks, L. Han, and Y. Guo, "Mining diverse small RNA species in the deep transcriptome," *Trends in Biochemical Sciences*, vol. 40, no. 1, pp. 4–7, 2015.

[77] X. Chen, Y. Ba, L. Ma et al., "Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases," *Cell Research*, vol. 18, no. 10, pp. 997–1006, 2008.

[78] M. Jung, A. Schaefer, I. Steiner et al., "Robust MicroRNA stability in degraded RNA preparations from human tissue and cell samples," *Clinical Chemistry*, vol. 56, no. 6, pp. 998–1006, 2010.

[79] P. S. Mitchell, R. K. Parkin, E. M. Kroh et al., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10513–10518, 2008.

[80] Y. Kakimoto, M. Tanaka, H. Kamiguchi, E. Ochiai, M. Osawa, and A. van Wijnen, "MicroRNA stability in FFPE tissue samples: dependence on GC content," *PLoS ONE*, vol. 11, no. 9, Article ID e0163125, 2016.

[81] W. Meng, J. P. McElroy, S. Volinia et al., "Comparison of microRNA deep sequencing of matched formalin-fixed paraffin-embedded and fresh frozen cancer tissues," *PLoS ONE*, vol. 8, no. 5, Article ID e64393, 2013.

[82] S. Tam, R. De Borja, M.-S. Tsao, and J. D. Mcpherson, "Robust global microRNA expression profiling using next-generation sequencing technologies," *Laboratory Investigation*, vol. 94, no. 3, pp. 350–358, 2014.

[83] L. Weng, X. Wu, H. Gao et al., "MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens," *The Journal of Pathology*, vol. 222, no. 1, pp. 41–51, 2010.

[84] J. L. Plieskatt, G. Rinaldi, Y. Feng et al., "Methods and matrices: approaches to identifying miRNAs for Nasopharyngeal carcinoma," *Journal of Translational Medicine*, vol. 12, no. 1, article 3, 2014.

[85] S. M. Riester, J. Torres-Mora, A. Dudakovic et al., "Hypoxia-related microRNA-210 is a diagnostic marker for discriminating osteoblastoma and osteosarcoma," *Journal of Orthopaedic Research*, 2016.

[86] R. De Paoli-Iseppi, P. A. Johansson, A. M. Menzies et al., "Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making," *Pathology*, vol. 48, no. 3, pp. 261–266, 2016.

[87] T. Holley, E. Lenkiewicz, L. Evers et al., "Deep clonal profiling of formalin fixed paraffin embedded clinical samples," *PLoS ONE*, vol. 7, no. 11, Article ID e50586, 2012.

[88] M. Kerick, M. Isau, B. Timmermann et al., "Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity," *BMC Medical Genomics*, vol. 4, article no. 68, 2011.

[89] S. Munchel, Y. Hoang, Y. Zhao et al., "Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics," *Oncotarget*, vol. 6, no. 28, pp. 25943–25961, 2015.

[90] E. Oh, Y.-L. Choi, M. J. Kwon et al., "Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples," *PLoS ONE*, vol. 10, no. 12, Article ID e0144162, 2015.

[91] S. Q. Wong, J. Li, R. Salemi et al., "Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours," *Scientific Reports*, vol. 3, article no. 3494, 2013.

[92] A. Mafficini, E. Amato, M. Fassan et al., "Reporting tumor molecular heterogeneity in histopathological diagnosis," *PLoS ONE*, vol. 9, no. 8, Article ID e104979, 2014.

[93] R. Bourgon, S. Lu, Y. Yan et al., "High-throughput detection of clinically relevant mutations in archived tumor samples by multiplexed PCR and next-generation sequencing," *Clinical Cancer Research*, vol. 20, no. 8, pp. 2080–2091, 2014.

[94] J. Ahn, K. S. Han, J. H. Heo et al., "FOXC2 and CLIP4: a potential biomarker for synchronous metastasis of ≤7-cm clear cell renal cell carcinomas," *Oncotarget*, vol. 7, no. 32, 2016.

[95] X. Castells, S. Karanović, M. Ardin et al., "Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid," *Cancer Epidemiology Biomarkers & Prevention*, vol. 24, no. 12, pp. 1873–1881, 2015.

[96] A. Collazo-Lorduy, M. Castillo-Martin, L. Wang et al., "Urachal carcinoma shares genomic alterations with colorectal carcinoma and may respond to epidermal growth factor inhibition," *European Urology*, vol. 70, no. 5, pp. 771–775, 2016.

[97] B. Jelakovic, X. Castells, K. Tomic, M. Ardin, S. Karanovic, and J. Zavadil, "Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid," *International Journal of Cancer*, vol. 136, no. 12, pp. 2967–2972, 2015.

[98] R. Menon, M. Deng, K. Rüenauver et al., "Somatic copy number alterations by whole-exome sequencing implicates *YWHAZ*

and *PTK2* in castration-resistant prostate cancer," *The Journal of Pathology*, vol. 231, no. 4, pp. 505–516, 2013.

[99] E. M. Van Allen, N. Wagle, P. Stojanov et al., "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine," *Nature Medicine*, vol. 20, no. 6, pp. 682–688, 2014.

[100] M. Kriegsmann, V. Endris, T. Wolf et al., "Mutational profiles in triple-negative breast cancer defined by ultradeep multigene sequencing show high rates of PI3K pathway alterations and clinically relevant entity subgroup specific differences," *Oncotarget*, vol. 5, no. 20, pp. 9952–9965, 2014.

[101] F. Meric-Bernstam, G. M. Frampton, J. Ferrer-Lozano et al., "Concordance of genomic alterations between primary and recurrent breast cancer," *Molecular Cancer Therapeutics*, vol. 13, no. 5, pp. 1382–1389, 2014.

[102] J. S. Ross, K. Wang, J. V. Rand et al., "Next-generation sequencing of adrenocortical carcinoma reveals new routes to targeted therapies," *Journal of Clinical Pathology*, vol. 67, no. 11, pp. 968–973, 2014.

[103] N. Wagle, M. F. Berger, M. J. Davis et al., "High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing," *Cancer Discovery*, vol. 2, no. 1, pp. 82–93, 2012.

[104] D. Korbie, E. Lin, D. Wall et al., "Multiplex bisulfite PCR resequencing of clinical FFPE DNA," *Clinical Epigenetics*, vol. 7, no. 1, article no. 28, 2015.

[105] M. Fang, L. Hutchinson, A. Deng, and M. R. Green, "Common BRAF(V600E)-directed pathway mediates widespread epigenetic silencing in colorectal cancer and melanoma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 5, pp. 1250–1255, 2016.

[106] H. Gu, C. Bock, T. S. Mikkelsen et al., "Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution," *Nature Methods*, vol. 7, no. 2, pp. 133–136, 2010.

[107] D. McCarthy, W. Pulverer, A. Weinhaeusel et al., "MethylMeter®: bisulfite-free quantitative and sensitive DNA methylation profiling and mutation detection in FFPE samples," *Epigenomics*, vol. 8, no. 6, pp. 747–765, 2016.

[108] R. P. Darst, C. E. Pardo, L. Ai, K. D. Brown, and M. P. Kladde, "Bisulfite sequencing of DNA," *Current Protocols in Molecular Biology*, chapter 7, Unit 7.9, pp. 1–17, 2010.

[109] M. Fanelli, S. Amatori, I. Barozzi et al., "Pathology tissue-chromatin immunoprecipitation, coupled with high-throughput sequencing, allows the epigenetic profiling of patient samples," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 50, pp. 21535–21540, 2010.

[110] P. Cejas, L. Li, N. K. O'Neill et al., "Chromatin immunoprecipitation from fixed clinical tissues reveals tumor-specific enhancer profiles," *Nature Medicine*, vol. 22, no. 6, pp. 685–691, 2016.

[111] G. K. Geiss, R. E. Bumgarner, B. Birditt et al., "Direct multiplexed measurement of gene expression with color-coded probe pairs," *Nature Biotechnology*, vol. 26, no. 3, pp. 317–325, 2008.

[112] B. Adam, B. Afzali, K. M. Dominy et al., "Multiplexed color-coded probe-based gene expression assessment for clinical molecular diagnostics in formalin-fixed paraffin-embedded human renal allograft tissue," *Clinical Transplantation*, vol. 30, no. 3, pp. 295–305, 2016.

[113] C. P. Kolbert, R. M. Feddersen, F. Rakhshan et al., "Multi-platform analysis of microRNA expression measurements in RNA from fresh frozen and FFPE tissues," *PLoS ONE*, vol. 8, no. 1, Article ID e52517, 2013.

[114] P. P. Reis, L. Waldron, R. S. Goswami et al., "mRNA transcript quantification in archival samples using multiplexed, color-coded probes," *BMC Biotechnology*, vol. 11, no. 1, article 46, 2011.

[115] A. Chatterjee, A. L. Leichter, V. Fan et al., "A cross comparison of technologies for the detection of microRNAs in clinical FFPE samples of hepatoblastoma patients," *Scientific Reports*, vol. 5, Article ID 10438, 2015.

[116] X. Chen, N. G. Deane, K. B. Lewis et al., "Comparison of nanostring nCounter® data on FFPE colon cancer samples and affymetrix microarray data on matched frozen tissues," *PLoS ONE*, vol. 11, no. 5, Article ID e0153784, 2016.

[117] J. Zhu, N. G. Deane, K. B. Lewis et al., "Evaluation of frozen tissue-derived prognostic gene expression signatures in FFPE colorectal cancer samples," *Scientific Reports*, vol. 6, p. 33273, 2016.

[118] M. H. Veldman-Jones, Z. Lai, M. Wappett et al., "Reproducible, quantitative, and flexible molecular subtyping of clinical DLBCL samples using the NanoString nCounter system," *Clinical Cancer Research*, vol. 21, no. 10, pp. 2367–2378, 2015.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2016, Article ID 6762731, 14 pages
http://dx.doi.org/10.1155/2016/6762731

Research Article

Comparative Transcriptome Analysis Reveals Effects of Exogenous Hematin on Anthocyanin Biosynthesis during Strawberry Fruit Ripening

Yi Li,[1,2] Huayin Li,[1] Fengde Wang,[1] Jingjuan Li,[1] Yihui Zhang,[1] Liangju Wang,[2] and Jianwei Gao[1]

[1]*Institute of Vegetables and Flowers, Shandong Academy of Agricultural Sciences and Shandong Key Laboratory of Greenhouse Vegetable Biology and Shandong Branch of National Vegetable Improvement Center, Jinan 250100, China*
[2]*College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China*

Correspondence should be addressed to Liangju Wang; wangliangju0909@sohu.com and Jianwei Gao; jianweigao3@qq.com

Received 15 July 2016; Revised 30 September 2016; Accepted 25 October 2016

Academic Editor: Quanhu Sheng

Anthocyanin in strawberries has a positive effect on fruit coloration. In this study, the role of exogenous hematin on anthocyanin biosynthesis was investigated. Our result showed that the white stage of strawberries treated with exogenous hematin had higher anthocyanin content, compared to the control group. Among all treatments, 5 μM of hematin was the optimal condition to promote color development. In order to explore the molecular mechanism of fruit coloring regulated by hematin, transcriptomes in the hematin- and non-hematin-treated fruit were analyzed. A large number of differentially expressed genes (DEGs) were identified in regulating anthocyanin synthesis, including the DEGs involved in anthocyanin biosynthesis, hormone signaling transduction, phytochrome signaling, starch and sucrose degradation, and transcriptional pathways. These regulatory networks may play an important role in regulating the color process of strawberries treated with hematin. In summary, exogenous hematin could promote fruit coloring by increasing anthocyanin content in the white stage of strawberries. Furthermore, transcriptome analysis suggests that hematin-promoted fruit coloring occurs through multiple related metabolic pathways, which provides valuable information for regulating fruit color via anthocyanin biosynthesis in strawberries.

1. Introduction

Strawberry (*Fragaria* × *ananassa* Duch.) is one of the most popular fruits with global economic importance [1]. Because of its appealing red coloration and abundant nutrition, strawberries are highly sought after by consumers [2, 3]. These qualities are partially due to the high anthocyanin content in strawberry. Anthocyanins have a high antioxidant activity [4]. Research suggests that anthocyanins have potential health benefits for a variety of conditions including cardiovascular disorders, advanced age-induced oxidative stress, inflammatory response [5], and diverse degenerative diseases [6, 7]. Increasing anthocyanin content in strawberries has been a relevant research topic in recent years.

Biosynthesis of anthocyanins is a complex biological process which is affected by genetic, developmental, and environmental factors [8]. Over the past few years, most structural genes encoding enzymes in the anthocyanin biosynthetic pathways have been isolated and characterized in strawberries. The first group of structural genes involved in these pathways includes phenylalanine ammonia lyase (PAL), chalcone synthase (CHS), flavanone 3-hydroxylase (F3H), dihydroflavonol-4-reductase (DFR), leucoanthocyanidin reductase (LAR)/anthocyanidin synthase (ANS), and UDP-glucose flavonoid 3-O-glucosyltransferase (UFGT). These structural genes comprise the pathways needed for the synthesis of anthocyanins [9]. These genes are regulated primarily by the Myb/bHLH/WD40 (MBW) complex in many plants [10, 11].

It has been shown that anthocyanin synthesis increases rapidly after the white stage in strawberries [12]. Phytohormones such as abscisic acid (ABA), cytokinin (CTK), and

ethylene and methyl jasmonate (JA) also play an important role in regulating the color development process in strawberries by increasing anthocyanin accumulation [12, 13]. Auxin and gibberellins (GAs) are known to reduce anthocyanin biosynthesis during the color development in fruit [8].

Light is one of the most important environmental factors regulating anthocyanin biosynthesis [14]. Phytochromes, which act as photoreceptors, play an important role in light stimulation during the development of strawberries [15]. Phytochromes are homodimeric chromoproteins where each holophytochrome is composed of a phytochrome protein covalently bound to a linear tetrapyrrole chromophore phytochromobilin (PΦB). PΦB acts as a light-receiving antenna for phytochrome. PΦB is synthesized in the plastid from heme catalyzed by heme oxygenase (HO) and subsequently phytochromobilin synthase [16]. Heme oxygenase 1 (HO1) is crucial to this process and acts as a rate-limiting enzyme in the biosynthesis of PΦB. HO1 catalyzes the oxygenation of heme to carbon monoxide, Fe^{2+}, and biliverdin (BV) in plants [17]. Additionally, HO1 has been shown to play an important role in anthocyanin accumulation in plants. For example, tomato [18, 19] and *Arabidopsis* [20–22] *HO1* deficient mutants, which can not synthesize the phytochrome chromophore, have a major reduction in anthocyanin accumulation.

Hematin ($C_{34}H_{33}O_5N_4Fe$), a protoporphyrin complex, is an inducer and substrate of HO1 in animals and plants [17, 23]. Exogenous hematin was shown to alleviate mercury-induced oxidative damage in the roots of *Medicago sativa* [17], induce adventitious root numbers and root length of cucumbers [24], regulate *Brassica nigra* seed germination under nanosilver stress, and relieve etiolation in the leaves of wheat seedlings under complete darkness [25, 26]. These effects might be derived from a hematin induced HO1 enzymatic reaction product.

In this study, the effect of exogenous hematin on the white stage of strawberries was investigated by comparing the anthocyanin contents in the hematin-treated fruit and the control group. In order to gain insight into the underlying molecular mechanisms regulating fruit coloring in response to the hematin treatment, an analysis of mRNA expression profiles was performed using high-throughput sequencing. The results demonstrated that exogenous hematin could promote fruit coloring. Comparative transcriptome analyses may give us a better understanding of the mechanism of the coloring process in strawberry fruit.

2. Materials and Methods

2.1. Plant Materials, Growth Condition, and Hematin Treatments. Strawberry (*Fragaria* × *ananassa* Duch cv. Benihoppe), an octoploid ($2n = 8X = 56$) species, was planted under standard culture conditions in a greenhouse (30/15°C, 14/10 h day/night, relative humidity 50–80%). The maximum light intensity inside the greenhouse was 55,300 lux. The developmental stage of the strawberry fruit was divided into seven visual stages: small green (SG), big green (BG), degreening (DG), white (Wt), initial red (IR), partial red (PR), and full red (FR) [12]. Strawberry plants at the white stage (about 25 d after anthesis) were chosen to study the effect of hematin on fruit coloration in the study. To study the effect of different concentration of hematin on fruit coloration, white stage strawberry plants ($n \geq 50$ for each treatment) were sprayed with 0, 1, 10, or 100 μM hematin (H3281, Sigma-Aldrich, St. Louis, MO, USA), respectively [26]. Strawberry fruits ($n \geq 30$ for each treatment) were harvested when the fruit entered the PR stage (48 h after treatment). The fruits were immediately frozen in liquid nitrogen and stored at −80°C for further analysis. We found the 10 μM hematin-treated strawberries accumulated most anthocyanin among all treatments (Figure S1, in Supplementary Material available online at http://dx.doi.org/10.1155/2016/6762731). By using the same method to treat the fruits with 0, 5, 10, or 15-μM hematin, we found the treatment by 5 μM hematin was the optimal condition for increasing anthocyanin production in strawberry (Figure 1) ($p < 0.01$). The treatment with 5 μM hematin was used for subsequent analysis. Three independent replicates were prepared for each treatment.

2.2. Determination of the Anthocyanin Content. The total anthocyanin content of the strawberries was determined using the method previously reported [6]. Absorbance was recorded on a Beckman DU640B spectrophotometer (Fullerton, CA, USA) at 510 and 700 nm for anthocyanin solutions in a pH 1.0 and pH 4.5 buffer, respectively. The calculated absorbance was obtained according to

$$A = \left[\left(A_{510} - A_{700}\right)_{\text{pH 1.0}} - \left(A_{510} - A_{700}\right)_{\text{pH 4.5}}\right]. \quad (1)$$

The molar extinction coefficient is 26,900 as described in other studies (e.g., [6]). Anthocyanin concentrations were expressed in milligrams of cyanidin-3-galactoside equivalent per gram of fresh weight. Three independent replicates were conducted for each treatment.

2.3. BV Preparation and Assay. Strawberry fruit (1 g) was homogenized in a Potter-Elvehjem homogenizer using 1.2 mL ice-cold 0.25 M sucrose solution containing 1 mM phenylmethyl sulfonyl fluoride, 0.2 mM EDTA, and 50 mM potassium phosphate buffer (pH 7.4). Homogenates were centrifuged at 20,000 ×g for 20 min and chloroplasts were used for activity determination. BV was assayed as previously described [27]. The concentration of BV was estimated using a molar absorption coefficient at 650 nm of 6.25 mM^{-1} cm^{-1} in 0.1 M HEPES-NaOH buffer (pH 7.2). Three independent replicates were conducted for each treatment.

2.4. RNA Isolation and cDNA Library Construction. Total RNA isolation was carried out as previously described [28]. The procedure is briefly presented below. Strawberry fruits were ground into powder and mixed at a ratio of 0.5 g powder to 20 mL extraction buffer (200 mM Tris-HCl (pH 8.2), 100 mM LiCl, 50 mM ETA, 1.5% SDS, 2% PVP (Sigma, PCP40), 2% BSA (Sigma), and 10 mM DTT (Sigma)). A total of 200 μL 10 mg/mL proteinase K (Merck, Darmstadt, Germany) was added to remove contaminating proteins. Total RNA was extracted using phenol/chloroform/isoamyl alcohol (25 : 24 : 1) and precipitated in a sodium acetate

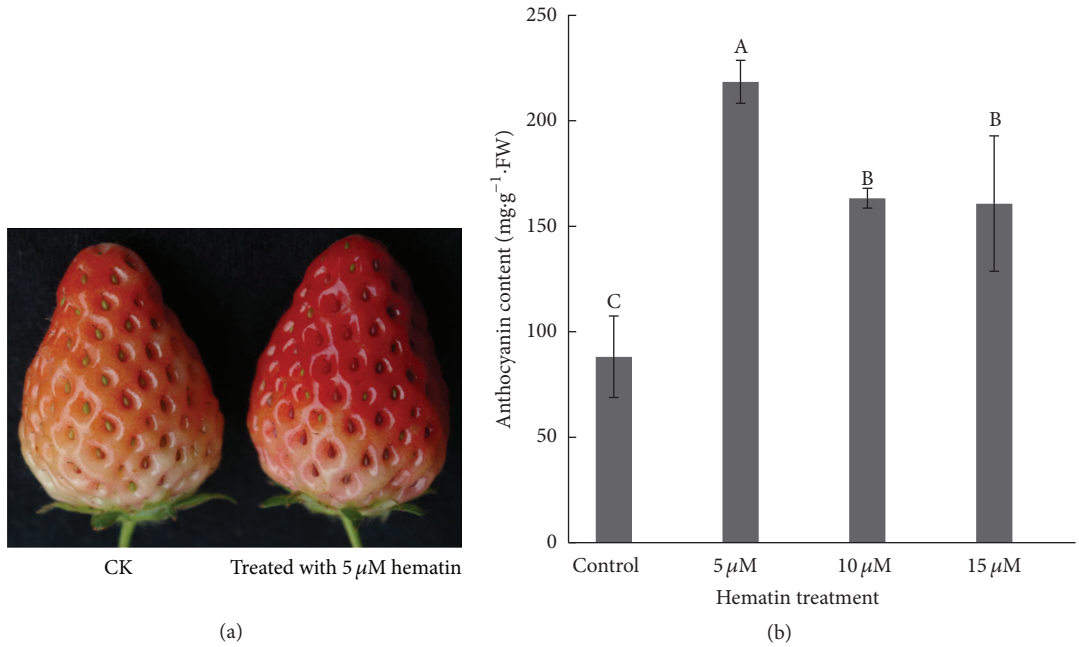


Figure 1: Fruit coloration and the anthocyanin content in the strawberry fruit. The color of the 5 μM hematin-treated strawberry fruit was redder than the control (CK) at 48 h after treatment (a). The anthocyanin content in the strawberry fruit was measured after treatment with different concentrations of hematin (0, 5, 10, and 15 μM) (b). Vertical bars represent standard errors; values with different letter are significantly different at $p < 0.01$.

and ethanol mixture. The mixture was resuspended in an appropriate volume of DEPC-treated distilled water and then stored at −80°C for the next step. The quality and quantity of the total RNA were measured using a NanoDrop ND-1000 Spectrophotometer (NanoDrop, Wilmington, DE, USA). Only samples that met the criteria of 1.8 ≤ OD260/280 ≤ D2.0 and OD260/230 ≥ 1.8 and concentration ≥ 200 ng/μL were used for sequencing. RNA samples of 30 strawberry fruits harvested in the same treatment group were pooled together for subsequent experiments.

RNA samples from two biological replicates were used for cDNA library construction and RNA-Seq at the Beijing Genomics Institute (BGI, Shenzhen, China). Total RNA samples were treated with DNase I (TaKaRa, Dalian, China) to remove any possible DNA contamination. The mRNA was enriched by using oligo (dT) magnetic beads (Illumina, San Diego, CA, USA) and cut into short fragments (about 200 bp). The first-strand cDNA was synthesized using a Superscript Preamplification System Kit (Gibco-BRL, Grand Island, NY, USA) as described in the manufacturer's instructions. The double stranded cDNA was purified with the oligo (dT) magnetic beads following the manufacturer's instructions. End repair was performed and adaptors were ligated to the ends of these fragments. Ligation products were purified using TAE-agarose gel electrophoresis. The fragments were then enriched by PCR amplification with an initial denaturing step at 98°C for 30 s, followed by 15 cycles of amplification (98°C for 10 s, 65°C for 30 s, and 72°C for 5 min) and a final extension at 72°C for 5 min. The PCR products were then purified using the oligo (dT) magnetic beads. DNA size, purity, and concentration were checked on an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA).

2.5. RNA Sequencing and Identification of Differentially Expressed Genes. RNA sequencing was performed using the ion proton platform at the Beijing Genomics Institute. The original sequence data were filtered to obtain clean reads for further analyses by removing short reads (less than 30 bp) and trimming adapters. Adapter reads were trimmed by first calculating the average quality of the first 15 bases from 3′-end until the average quality was larger than 10 and then removing the bases that were counted. The high-quality clean reads were mapped against the strawberry reference genome (http://strawberry-garden.kazusa.or.jp/) using Ion Torrent's mapping program (TMAP, version 0.2.3; https://github.com/iontorrent/TMAP). No more than two mismatches were allowed in the sequence alignment. Quality assessment of reads, statistics of alignment, sequencing saturation analysis, and randomness assessments were carried out subsequently to assess the quality of sequencing.

Gene expression levels were quantified by the software Sailfish [29]. Raw counts were normalized to Reads Per Kilobase of exon model per Million mapped reads (RPKM). Differential expression analysis was performed using the EBSeq package [30]. Q value < 0.05 and $|\log_2(\text{fold change})| > 1$ were set as the threshold to identify significant differentially expressed genes. In statistics, correction for false positive errors was carried out using the FDR statistic [31].

2.6. Gene Ontology and Pathway Enrichment Analysis of Differentially Expressed Genes. Gene functions of strawberry were annotated according to the Gene Ontology (GO) standardized terms for molecular function, cellular component, and biological process using Blast2GO (https://www.blast2go.com/). Gene function was annotated using the following databases: NCBI nonredundant protein sequences, NCBI nonredundant nucleotide sequences, protein family, Clusters of Orthologous Groups of proteins, the manually annotated and reviewed protein sequence database, KEGG Ortholog database, and Gene Ontology. After GO annotation for DEGs, we performed GO functional classification for DEGs by the WEGO software [32] and analyzed the distribution of gene functions. We used the GO Term Finder tool (http://www.yeastgenome.org/help/analyze/go-term-finder) to search for significant shared GO terms.

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}, \quad (2)$$

where N is the number of all genes with GO annotation; n is the number of DEGs in N; M is the number of all genes that are annotated to certain GO terms; and m is the number of DEGs in M. The calculated p value was subjected to Bonferroni Correction [33]. The corrected p value < 0.05 was set as the threshold. GO terms fulfilling this condition are defined as significantly enriched GO terms in DEGs. The pathway enrichment analyses using the KEGG database (http://www.genome.jp/kegg/) were conducted subsequently to study the functions of differentially expressed genes identified between the control and the hematin-treated strawberry groups. The formula is the same as that in GO analysis. Here N is the number of all genes with KEGG annotation, n is the number of DEGs in N, M is the number of all genes annotated to specific pathways, and m is the number of DEGs in M.

2.7. Quantitative Real-Time PCR Analysis. Quantitative real-time PCR (qRT-PCR) experiments were conducted to assess the reliability of the RNA-Seq data [34]. Eleven pairs of gene specific primers were designed to verify the DEGs in the strawberry fruit (Table S1). Strawberry RNA was extracted from the fruit according to the previously detailed method. Total RNA was digested with DNase I for 30 min at 25°C to remove DNA contamination according to the manufacturer's instructions. The qRT-PCR was performed using a SYBR PCR master mix (TaKaRa, Dalian, China) on a Bio-Rad IQ-5 thermal cycler (Bio-Rad, Philadelphia, PA, USA). Three replicates of each sample were conducted to calculate the average Ct values. The relative expression level was calculated by the comparative $2^{-\Delta\Delta Ct}$ method [35, 36]. The significance was determined with the SPSS software (SPSS 17.0, IBM, Chicago, IL, USA) ($p < 0.05$).

3. Results and Discussion

3.1. Effects of Exogenous Hematin on Anthocyanin and Biliverdin Accumulation. Hematin is an inducer and substrate of HO1 in animals and plants [17, 23]. In this study, strawberries at the Wt stage were treated with 0 μM (control), 5, 10, or 15 μM hematin. The results indicated that the anthocyanin content in the hematin-treated strawberries was more than the control after 48 h. The treatment by 5 μM hematin was found to be the optimal condition for increasing anthocyanin production (Figure 1) ($p < 0.01$). The anthocyanin content in the 5 μM hematin-treated strawberries was 2.5 times higher than that in the control. This is the first report that hematin could increase the anthocyanin production in fruit. We measured the expression of *FaHO-1* and the content of biliverdin (BV) which is the metabolite of heme oxygenase. We found that hematin could also significantly increase the expression of *FaHO-1* in strawberry fruit (Figure 4) and promote the accumulation of BV (Figure 2). These results indicate that hematin promotes anthocyanin accumulation through the heme metabolism pathway.

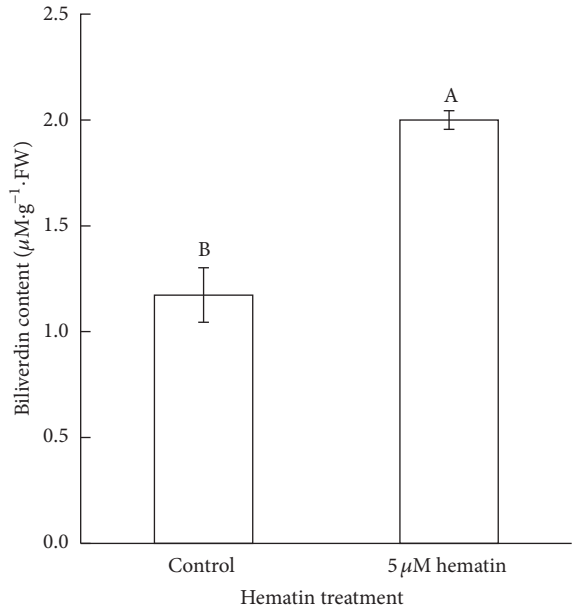


Figure 2: The BV content in strawberry fruit was measured after treatment with 0 and 5 μM hematin. Vertical bars represent standard errors; values with different letter are significantly different at $p < 0.01$.

3.2. Sequences Assembly, Mapping, and Functional Annotation. High-throughput sequencing technology is utilized widely in analyzing gene expression in many organisms [37, 38]. In this study, approximately 10.8 Gbp raw tags were generated for each library. After eliminating adapters, ambiguous nucleotides, and low-quality sequences, a total of 34,618,832 and 45,464,123 clean reads between 150 and 200 nucleotides in length were obtained (Table 1). Over 85% of the clean tags from each library mapped to reference genes with less than 2 bp mismatches. Less than 29% of the clean tags from each library could not be aligned to any reference genes because of incomplete sequences, and these tags were designated as unknown. More than 67.3% of the clean tags in each library were mapped to a single gene, while less than 4.7% mapped to multiple reference genes. Unknown tags and tags mapped to multigenes were filtered out, and the unique

Table 1: Overview of the sequencing and assembly.

	Biological replicate 1		*Biological replicate 2*	
Sample	Control	Hematin-treated	Control	Hematin-treated
Clean reads	17808924	26809908	19235452	26228671
Genome map rate (%)	96.58	96.52	96.65	96.74
Gene map rate (%)	86.46	86.27	87.04	85.63
Perfect match (%)	4677273 (24.32%)	6510668 (24.28%)	4414050 (24.79%)	6149508 (23.45%)
Unique match (%)	15075148 (84.65%)	22628939 (84.41%)	16409557 (85.31%)	21986962 (83.83%)
Expressed genes	27976	28713	28999	28211
Mismatch (%)	11321162 (63.57%)	17101590 (63.79%)	12456871 (64.76%)	16697280 (63.66%)
Total unmapped reads	645135 (3.35%)	932213 (3.48%)	608914 (3.42%)	856028 (3.26%)

clean tags that mapped to a single gene were retained for further DEG analysis.

In addition, the perfect match rates were 24.32%, 24.28%, 24.79%, and 24.35% in control 1, control 2, hematin 1, and hematin 2, respectively.

The saturation of the libraries with and without the hematin treatment was analyzed. The number of the detected genes became saturated at about 2 million reads (Figure S2). This indicates that the obtained reads are sufficient for complete transcriptome coverage. The randomness of RNA fragmentation in the four libraries was assessed for subsequent bioinformatics analysis. The results showed that the reads in each position of the reference gene were distributed evenly and demonstrated highly similar tendencies in all libraries (Figure S3).

3.3. Genes Differentially Regulated in Response to Exogenous Hematin. Based on sequencing results of the four mRNA libraries, approximately 29,000 genes (about 70% of the reference genes in the octoploid strawberry genome) were detected in each library. Additionally, correlations among genes based on RPKM between the two biological replicates were analyzed; the correlation coefficients (R^2) were high (0.93 and 0.95 for the control and hematin-treated group, resp.). Genes expressed in both replicates were screened. A total of 28,713, 28,999, 27,976, and 28,211 genes (Table 1) were expressed in the control and hematin-treated samples.

To reveal the molecular pathways regulating strawberry coloring in response to hematin, the DEGs between the control and hematin-treated fruit were analyzed. We compared the gene expression profiles between the control and the treated samples in both biological replicates. DEGs detected in both biological replicates were screened for subsequent analysis. A total of 1,080 (402 up- and 678 downregulated) genes were differentially expressed in the 5 μM hematin-treated groups compared with the control groups. More genes were downregulated in the hematin-treated group compared to the control group.

To facilitate the global analysis of gene expression, a GO analysis was performed by mapping each differentially expressed gene into the records of the GO database. Gene Ontology (GO) enrichment analysis of the DEGs in the control and hematin-treated groups was performed to reveal the possible mechanisms under which hematin promotes anthocyanin biosynthesis. GO terms with $p < 0.05$ were represented among genes with significant changes in expression over a given interval. In this study, only three terms including anatomical structure arrangement, meristem structural organization, radial pattern formation were significantly enriched in biological process (Supplementary Table S2). Although most Go terms were not significantly enriched, they could provide a reference for further study. According to the GO cellular components, a large number of the DEGs were classified into main cell organelles, such as vacuole, nucleolus, mitochondrion, apoplast, chloroplast, plastid, cytoplasm, and membrane. In particular, many DEGs were assigned to the plastid and chloroplast (Figure 3), indicating that hematin has an important impact on the expression of plastid and chloroplast genes. It is probably because heme oxygenase 1 is a soluble plastid protein [22]. According to the GO molecular function, a large number of the DEGs were classified into UDP-glucosyltransferase, sucrose transmembrane transporter activity, nucleic acid binding, and DNA binding terms which are closely related to anthocyanin biosynthesis. Many DEGs were classified into heme binding, peroxidase activity, and tetrapyrrole binding (Figure 3). It is probably because hematin can increase the activity of HO1 and promote the degradation of hemoglobin [17]. In addition, anthocyanin accumulation in fruit is closely related to hormone [8] and carbohydrate biosynthesis [9]. According to the GO biological process, a large number of the DEGs were classified into regulation of hormone levels, cytokinin metabolic process, response to hormone, response to abscisic acid, response to auxin, and response to ethylene terms, indicating that the hormone-related DEGs are involved in regulating the anthocyanin biosynthesis in the hematin-treated fruit. Many DEGs were classified into starch and sucrose biosynthetic and transport process and phenylpropanoid metabolic process terms, indicating that hematin can promote anthocyanin accumulation via the phenylpropanoid metabolic and carbohydrate metabolism. In addition, many DEGs were classified into the process for response to light stimulus term, indicating that hematin may also participate in the light stimulus system.

Pathway enrichment analysis was performed to understand the biological functions of the DEGs by identifying significantly enriched signal transduction pathways or metabolic pathways [39]. In the study, a number of altered biological pathways associated with the hematin treatment

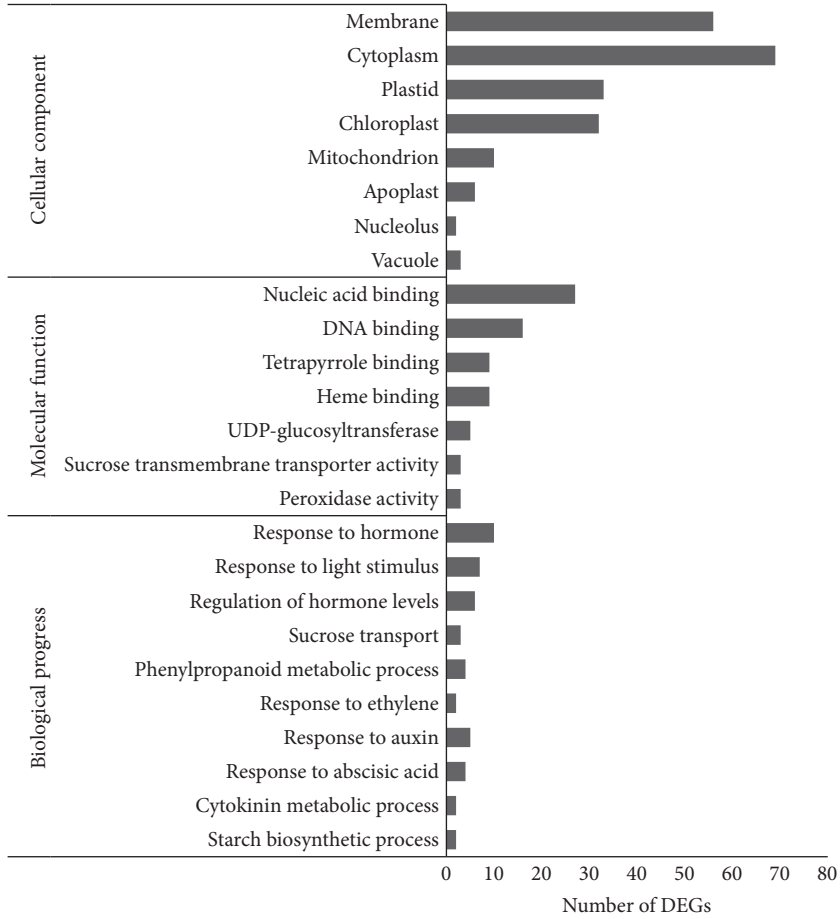


Figure 3: Gene classification based on Gene Ontology (GO) for DEGs in the strawberries treated by hematin.

were identified, and three pathways, including RNA polymerase, pyrimidine metabolism, and purine metabolism, were significantly enriched (Q-value < 0.05) (Table 2). Although the majority of the pathway terms were not significantly enriched, the pathway enrichment analysis helps us to further understand the biological functions of the DEGs and the molecular mechanisms that regulate fruit coloring in response to the hematin treatment. These pathway terms included metabolic pathways, biosynthesis of secondary metabolites, isoflavonoid biosynthesis, phenylalanine metabolism, phenylalanine, tyrosine and tryptophan biosynthesis, starch and sucrose metabolism, plant hormone signal transduction, flavone and flavonol biosynthesis, and phenylpropanoid biosynthesis terms (Supplementary Table S2).

3.4. Analysis of DEGs Involved in Anthocyanin Biosynthesis

3.4.1. Transcription Factors. Transcription factors (TFs) are proteins that regulate the expression of downstream target genes. Many TFs, such as v-myb avian myeloblastosis viral oncogene homolog (MYB), basic helix-loop-helix (bHLH), WD40-repeats protein (WD40), and MADS-box (MADS), directly regulating the expression of the structural genes in anthocyanin biosynthesis have been identified from many species [8]. Therefore, the enhanced expression of key TFs might regulate the anthocyanin biosynthesis. In our study, some anthocyanin biosynthesis-related TFs were found to be uniquely present in the DEG profiling of the hematin-treated fruit, for example, MYB, bHLH, PIF3, MADs-box, AP2-EREBP, and ABI3/VP1 (Table 3). The anthocyanin biosynthesis genes are regulated primarily by the MBW transcription factor, which is a ternary transcription factor complex [10, 11]. Most of the MYBs involved in regulating anthocyanin biosynthesis are positive regulators of transcription [8]. However, MYBs can act as repressors too, such as strawberry *FaMYB1* and *FaMYB9* and grapevine *VvMYB4,* which can significantly suppress the biosynthesis of anthocyanins and flavonols [40]. In this study, three unigenes belonging to MYB transcription factors were differentially expressed, and two MYB transcription factors were significantly upregulated. Of these, the unigene "FANhyb_rscf00000146.1.g00007.1" from the *MYB* family of R2R3 MYB transcription factors was significantly upregulated. Moreover, increasing evidence indicates that the expression of the bHLHs promotes anthocyanin accumulation in fruit [41]. In our study, all of the unigenes from the bHLH family were downregulated, among which the expression level of unigene FANhyb_rscf00001292.1.g00003.1 was significantly decreased 6250-fold (Table 3). FANhyb_rscf00002210.1.g00001.1 encodes a predicted MADS-box transcription factor, and it was

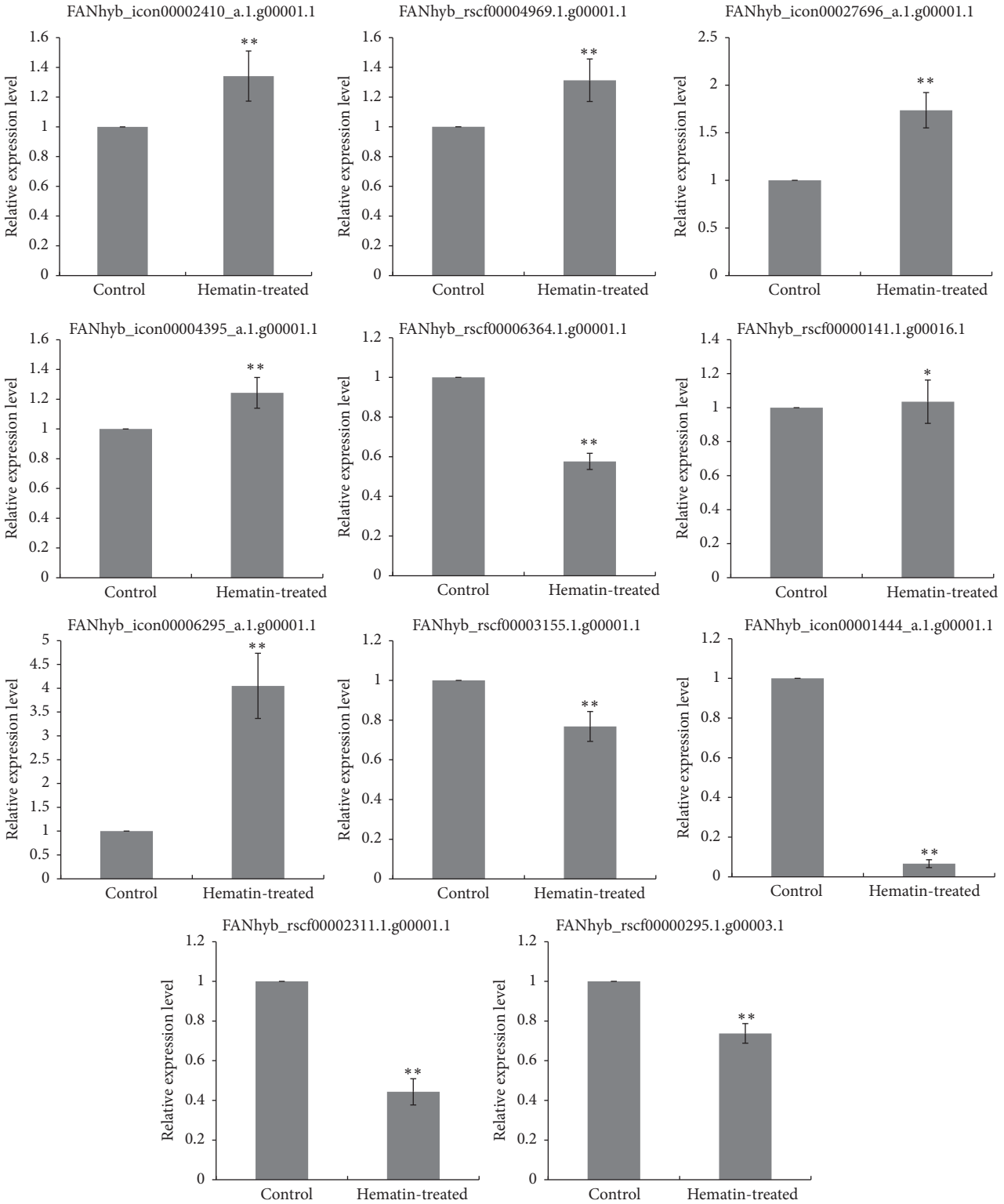


Figure 4: qRT-PCR validation of the RNA-Seq based gene expression. The values indicate means of three biological replicates ± SD. Star indicates that the expression level is significantly different between the hematin-treated and the control group ($^{*}p < 0.05$, $^{**}p < 0.01$).

Table 2: Pathway enrichment analysis of differentially expressed genes.

Pathway	DEGs with pathway annotation (451)	All genes with pathway annotation (15877)	Corrected p value	Q-value	Pathway ID
RNA polymerase	16	183	$6.88E-5$	0.0068107	ko03020
Pyrimidine metabolism	20	317	0.0007651746	0.0320	ko00240
Purine metabolism	21	347	0.0009710371	0.032044224	ko00230

significantly downregulated. There were three unigenes in the AP2-EREBP family, with only one upregulated and the others downregulated. These results suggest that exogenous hematin is involved in developmental transcriptional regulation of anthocyanin biosynthesis.

3.4.2. DEGs in Anthocyanin Biosynthesis. The anthocyanin biosynthetic pathway has been studied intensively. The anthocyanin biosynthetic pathway via the phenylpropanoid pathway is well known [8]. In this study, the transcriptome data showed a significant increase in the expression of genes in the anthocyanin biosynthesis pathway when exposed to exogenous hematin. The structural genes involved in anthocyanin biosynthesis, including DFR, LAR, and UDP-glycosyltransferase, exhibited significant differential expression in response to the hematin treatment (Table 4). For example, two *DFR* genes and one *LAR* gene were significantly upregulated (the expression increased 18-, 30-, and 3-fold in the hematin-treated fruit, resp.). In addition, all six genes encoding *UDP-glycosyltransferase* were significantly upregulated. Among these, the expression of one gene (Fanhy_icon15742070o.1.g00001.1) was upregulated 18305-fold. The results suggest that exogenous hematin may have an important role in promoting anthocyanin biosynthesis.

3.4.3. DEGs in Hormone Signaling Transduction in Response to Hematin. Auxin has been shown to negatively regulate the expression of the anthocyanin biosynthesis genes [42]. In this study, there were three unigenes involved in the auxin signaling pathway, and all were downregulated, including one gene encoding auxin efflux carrier component 1-like (AUX1-like) and two genes encoding auxin response factors (ARF). These results suggest that hematin promotes anthocyanin accumulation by regulating auxin in the coloration process of strawberries. It has shown that auxin suppresses anthocyanin biosynthesis in the red-fleshed apple callus [43]. Endogenous expression of auxin [44, 45] has been found to impede anthocyanin accumulation in strawberries. In additional, cytokinins are also known to play an important role in anthocyanin biosynthesis. It has shown to enhance anthocyanin accumulation in *Zea mays* and regulate anthocyanin production and composition in suspension cultures of strawberry cells [46]. In this study, two unigenes related to the cytokinin signaling pathway were identified, and they were downregulated as well. One gene was identified as cytokinin dehydrogenase 5-like which catalyzes the irreversible degradation of cytokinins. This result indicates that hematin promotes anthocyanin accumulation by regulating the cytokinin signaling pathway.

Extensive researches have shown that ABA plays an important role in the regulation of anthocyanin biosynthesis in nonclimacteric fruit [47]. In this study, a total of five unigenes related to the ABA signaling pathway were identified, among which two were upregulated and three were downregulated. Two genes, which are involved in the abscisic acid response, were upregulated. These two unigenes encode 9-cis-epoxycarotenoid dioxygenase 1 (NCED1) and ABA overly sensitive 5 (ABO5), respectively. ABA is synthesized from carotenoids via several enzymatic reactions in the plastid. The rate-limiting step in these reactions is catalyzed by 9-cis-epoxycarotenoid dioxygenase 1 (NCED1) [48]. Mutations in the *FaNCED1* gene result in colorless strawberry fruit, which can become colored through the application of exogenous ABA [12]. Another unigene was identified to encode abscisic acid 8′-hydroxylase 1-like, a key enzyme in the oxidative catabolism of abscisic acid, which was downregulated. Application of exogenous ABA regulates phenylalanine ammonia lyase activity and increases the phenolic and anthocyanin content of strawberry fruit [12, 49]. Our results suggest that hematin promotes ABA biosynthesis and inhibits ABA disintegration in the hematin-treated fruit.

In *Arabidopsis*, JAs can affect anthocyanin accumulation via the interaction of negative regulators with the MBW complex of transcription factors involved in anthocyanin biosynthesis [50]. Preharvest application of JA to "*Fuji*" apples enhances red coloration [51]. JA vapor treatment can also enhance anthocyanins in strawberry fruit [52]. In this study, jasmonate O-methyltransferase-like exhibited significantly differential expression. Jasmonate O-methyltransferase-like is also known to be a key enzyme for jasmonate-regulated plant responses to stimuli [53, 54]. These results suggest that hematin increases the activity of the JA-regulated anthocyanin biosynthesis.

3.4.4. DEGs in the Phytochrome Signaling Pathway in Response to Hematin. Phytochromes that act as photoreceptors play an important role in anthocyanin regulation [15]. Phytochromes are homodimeric chromoproteins, where each holophytochrome is composed of a phytochrome protein (apophytochrome) covalently bound to a linear tetrapyrrole PΦB. HO1 is crucial to this process and acts as a rate-limiting enzyme in the biosynthesis of PΦB [55]. Exogenous hematin can also induce *HO-1* expression in many plants [56, 57]. In this study, we found exogenous hematin significantly increased the expression of *FaHO1* (Table 5). This result implies that hematin can promote PΦB biosynthesis. Recently, it has been reported that hematin could induce the accumulation of far-red phytochrome and *phytochrome*

Table 3: DEGs acting as transcription factors in response to exogenous hematin.

Annotation transcription factors	Gene ID	Real fold-change values	Upregulation/downregulation	FDR
PREDICTED: AP2-like ethylene-responsive transcription factor AIL1-like	FANhyb_icon00006295_a.1.g00001.1	9.38	Up	$1.20E-05$
PREDICTED: AP2-like ethylene-responsive transcription factor AIL1-like	FANhyb_rscf00003155.1.g00001.1	4	Down	$4.80E-16$
PREDICTED: AP2-like ethylene-responsive transcription factor AIL5-like	FANhyb_rscf00000029.1.g00010.1	4	Down	$3.20E-08$
MYBdomain protein 66	FANhyb_rscf00000146.1.g00007.1	6.54	Up	$7.80E-07$
MYB-related protein 3R-1-like, partial	FANhyb_rscf00000141.1.g00016.1	2.82	Up	$8.30E-04$
PREDICTED: transcription factor GAMYB-like	FANhyb_rscf00000649.1.g00006.1	3.03	Down	$1.50E-13$
PREDICTED: probable WRKY transcription factor 53-like	FANhyb_rscf00001973.1.g00002.1	2.17	Down	$2.00E-08$
Transcription factor bHLH80-like	FANhyb_rscf00001292.1.g00003.1	6250	Down	$1.50E-06$
Transcription factor bHLH70-like	FANhyb_rscf00000295.1.g00003.1	5.26	Down	$6.60E-12$
Transcription factor ICE1-like	FANhyb_rscf00000170.1.g00006.1	2.44	Down	$4.50E-18$
C2H2-like zinc finger protein	FANhyb_rscf00001143.1.g00002.1	2.82	Up	$5.90E-04$
Zinc finger protein MAGPIE-like	FANhyb_rscf00003024.1.g00001.1	2.04	Up	$1.50E-04$
PREDICTED: axial regulator YABBY 5-like	FANhyb_rscf00001667.1.g00001.1	4.17	Down	$7.30E-05$
PREDICTED: axial regulator YABBY 1-like	FANhyb_rscf00006899.1.g00001.1	3.57	Down	$6.20E-06$
PREDICTED: dof zinc finger protein DOF3.7-like isoform 1	FANhyb_rscf00006856.1.g00001.1	2.63	Down	$2.80E-04$
ZF-HD homeobox protein	FANhyb_rscf00000044.1.g00022.1	3.85	Down	$4.20E-08$
ZF-HD homeobox protein At4g24660-like	FANhyb_rscf00005241.1.g00001.1	3.57	Down	$4.70E-09$
PREDICTED: B3 domain-containing protein REM14-like	FANhyb_rscf00001009.1.g00002.1	2.43	Up	$5.30E-04$
PREDICTED: B3 domain-containing transcription factor ABI3-like	FANhyb_rscf00001271.1.g00002.1	4.55	Down	$1.30E-28$
Growth-regulating factor 6	FANhyb_rscf00000024.1.g00029.1	3.45	Down	$3.30E-34$
Growth-regulating factor 5	FANhyb_icon00009704_a.1.g00001.1	2.7	Down	$1.90E-06$
GATA type zinc finger transcription factor family protein	FANhyb_rscf00000393.1.g00013.1	11.11	Down	$1.10E-05$
PREDICTED: zinc finger CCCH domain-containing protein 2-like	FANhyb_icon00006770_a.1.g00001.1	3.13	Down	$6.80E-49$
PREDICTED: E2F transcription factor-like E2FE-like	FANhyb_icon00002702_a.1.g00001.1	2.78	Down	$2.40E-04$
PREDICTED: uncharacterized protein	FANhyb_rscf00001066.1.g00001.1	2.63	Down	$4.90E-06$
PREDICTED: nuclear transcription factor Y subunit A-3-like isoform 1	FANhyb_rscf00000008.1.g00035.1	3.85	Down	$5.00E-20$
PREDICTED: homeobox-leucine zipper protein ATHB-8-like	FANhyb_rscf00000015.1.g00005.1	2.63	Down	$1.00E-06$
PREDICTED: MADS-box transcription factor 18-like	FANhyb_rscf00000323.1.g00016.1	2.43	Down	$1.80E-06$
PREDICTED: zinc finger protein 132-like isoform 1	FANhyb_rscf00007015.1.g00001.1	2.43	Down	$1.80E-08$
PREDICTED: transcription factor PIF3-like	FANhyb_rscf00000669.1.g00002.1	3.46	Up	$1.40E-07$

Table 4: Regulation of the DEGs in anthocyanin biosynthesis in response to exogenous hematin.

Genes in anthocyanin biosynthesis	Gene ID	Real fold-change values	Regulation level	FDR
Dihydroflavonol-4-reductase	FANhyb_icon00002410_a.1.g00001.1	30.91	Up	$9.10E-06$
Flavonol synthase	FANhyb_icon00015354_a.1.g00001.1	3.45	down	$1.80E-05$
Isoflavone reductase homolog	FANhyb_icon20341135_s.1.g00001.1	3.34	Up	$7.50E-14$
UDP-glycosyltransferase	FANhyb_rscf00004969.1.g00001.1	3.07	Up	$6.40E-08$
UDP-glycosyltransferase activity	FANhyb_icon15742070_o.1.g00001.1	18 305.63	Up	$5.10E-09$
UDP-glycosyltransferase 73C1	FANhyb_icon00027696_a.1.g00001.1	2.98	Up	$5.60E-04$
UDP-glycosyltransferase activity	FANhyb_icon00008064_a.1.g00001.1	2.49	Up	$3.10E-04$
PREDICTED: UDP-glucose flavonoid 3-O-glucosyltransferase 7-like	FANhyb_rscf00005563.1.g00001.1	2.12	Up	$1.10E-19$
PREDICTED: UDP-glycosyltransferase 76F1-like	FANhyb_rscf00002177.1.g00003.1	2	Up	$7.20E-29$

A transcripts in the etiolated leaves of wheat seedling phytochromes [26]. However, our study shows that phytochrome A was not differentially expressed. It is probably because the hematin-treated strawberries were not shaded during the experiment, while the function of phytochrome A is light-dependent. In addition, we found that a downstream gene of phytochrome, PIF3 (FANhyb_rscf00000669.1.g00002.1), was significantly upregulated. PIF3 is thought to be a positive regulator of phytochrome B mediated by light-dependent signal transduction [58]. These results suggest that the phytochrome pathways may be involved in the anthocyanin biosynthesis promoted by hematin.

3.4.5. DEGs in Starch and Sucrose Degradation. Proteomic approaches have revealed that starch degradation contributes to anthocyanin accumulation in tuberous roots of the purple sweet potato variety [59]. Starch and sucrose biosynthesis is important for anthocyanin accumulation in strawberry fruit [60]. In this study, we found the DEGs involved in starch and sucrose synthesis in the hematin-treated fruit were all downregulated (Table 5). For example, the expression of the unigenes encoding soluble starch synthase 3, sucrose transport protein SUC2-like, and a probable sucrose-phosphate synthase 4-like enzyme were decreased 2.3-, 4.6-, 3.2-, and 2-fold, respectively. This suggests that exogenous hematin regulates the biosynthesis of starch and sucrose and hence affects fruit coloring.

3.4.6. DEGs in the Calcium Pathway in Response to Hematin. Calcium can also increase the transcription levels of key structural genes *F3H*, *DFR*, *ANT,* and *UFGT* in the white stage of strawberries [61]. In this study, we found that three DEGs in calcium biosynthesis and transport were downregulated in the hematin-treated fruit (Table 5). Only the unigene encoding predicted cation/calcium exchanger 5-like was upregulated. The unigenes encoding a calpain-type cysteine protease and a predicted calcium-binding protein PBP1-like were downregulated, respectively. Exogenous application of calcium can promote apple coloring [62] in addition to variation in anthocyanin content [61]. Our results indicate that the biosynthesis and transport of calcium are involved in the development of coloring in hematin-treated strawberries.

3.5. Validation of Selected DEGs by qRT-PCR. To validate the expression of the DEGs obtained from RNA-Seq, 11 DEGs were selected for qRT-PCR, including structural genes (*DFR* and *UDFGs*), transcription factor genes (*MYB* and *bHLH*), and phytochrome chromophore-related gene (*HO-1*). The primers used for qRT-PCR are listed in Supplementary Table S1. The qRT-PCR results were consistent with the RNA-Seq data (Figure 4), except for FANhyb_rscf00000141.1.g00016.1, which had a higher $\log_2$ ratio (hematin-treated/control) in the transcriptome data than in qRT-PCR. These results indicate the RNA-Seq data from the strawberry transcriptome is reproducible and accurate.

4. Conclusions

In this study, the anthocyanin content in the strawberry fruit was elevated by the application of exogenous hematin. This is the first report that hematin could increase the anthocyanin production in fruit. Furthermore, we explored the effects of the exogenous hematin on metabolic pathways using genome-wide transcriptome analysis. The results indicate that the expression levels of many genes involved

Table 5: Other DEGs identified in anthocyanin biosynthesis-related pathways.

Anthocyanin biosynthesis-related pathways	Annotation genes	Gene ID	Real fold-change values	Upregulation/downregulation	FDR
Calcium ion binding	Cation/calcium exchanger 5-like	FANhyb_rscf00001906.1.g00001.1	2.03	Up	$6.4E-06$
	Calpain-type cysteine protease family	FANhyb_icon00011755_a.1.g00001.1	3.45	Down	$1.8E-09$
	Calcium-binding protein PBP1-like	FANhyb_rscf00000750.1.g00008.1	3.23	Down	$9.7E-04$
	Calpain-type cysteine protease family	FANhyb_icon00023658_a.1.g00001.1	2.33	Down	$6.0E-08$
Cytokinin	Cytokinin dehydrogenase 5-like	FANhyb_iscf00325393_1_s.1.g00001.1	6.67	Down	$6.2E-06$
	Cytochrome P450 714A1	FANhyb_rscf00000592.1.g00003.1	2.17	Down	$9.8E-10$
Jasmonate	Jasmonate O-methyltransferase-like	FANhyb_rscf00000004.1.g00013.1	2.01	Up	$7.0E-04$
Abscisic acid	Protein ABSCISIC ACID-INSENSITIVE 5-like	FANhyb_rscf00002164.1.g00001.1	4	Down	$1.6E-18$
	ABA overly sensitive 5	FANhyb_icon00051144_a.1.g00001.1	2.91	Up	$2.4E-05$
	9-cis-epoxycarotenoid dioxygenase NCED1	FANhyb_icon18399909_o.1.g00001.1	4.41	Up	$4.3E-06$
	Abscisic acid 8′-hydroxylase 1-like	FANhyb_icon00000938_a.1.g00001.1	2.56	Down	$1.4E-06$
	Abscisic acid receptor PYR1-like	FANhyb_icon00020426_a.1.g00001.1	2.08	Down	$3.0E-06$
Auxin	Auxin efflux carrier component 1-like	FANhyb_rscf00001008.1.g00003.1	3.84	Down	$2.3E-06$
	Auxin response factor 8	FANhyb_rscf00001306.1.g00002.1	2.63	Down	$6.1E-17$
	Auxin response factor 17-like	FANhyb_rscf00006074.1.g00001.1	2.17	Down	$1.1E-36$
Starch and sucrose	Glycosyltransferase, family 35	FANhyb_rscf00000034.1.g00008.1	3.84	Down	$1.1E-05$
	Soluble starch synthase 3	FANhyb_rscf00000045.1.g00004.1	2.32	Down	$3.8E-12$
	Sucrose transport protein SUC2-like	FANhyb_icon00036727_a.1.g00001.1	4.55	Down	$9.8E-04$
	Sucrose transport protein SUC2-like	FANhyb_rscf00000755.1.g00002.1	3.23	Down	$6.6E-10$
	Probable sucrose-phosphate synthase 4-like	FANhyb_rscf00001350.1.g00002.1	2.13	Down	$1.1E-14$
Phytochrome	PREDICTED: transcription factor PIF3-like	FANhyb_rscf00000669.1.g00002.1	3.46	up	$1.4E-07$
	Phytochrome E-like	FANhyb_rscf00000436.1.g00004.1	4	Down	$3.3E-05$
	Heme oxygenase 1	FANhyb_icon00004395_a.1.g00001.1	2.58	$2.7E-07$	$2.7E-07$

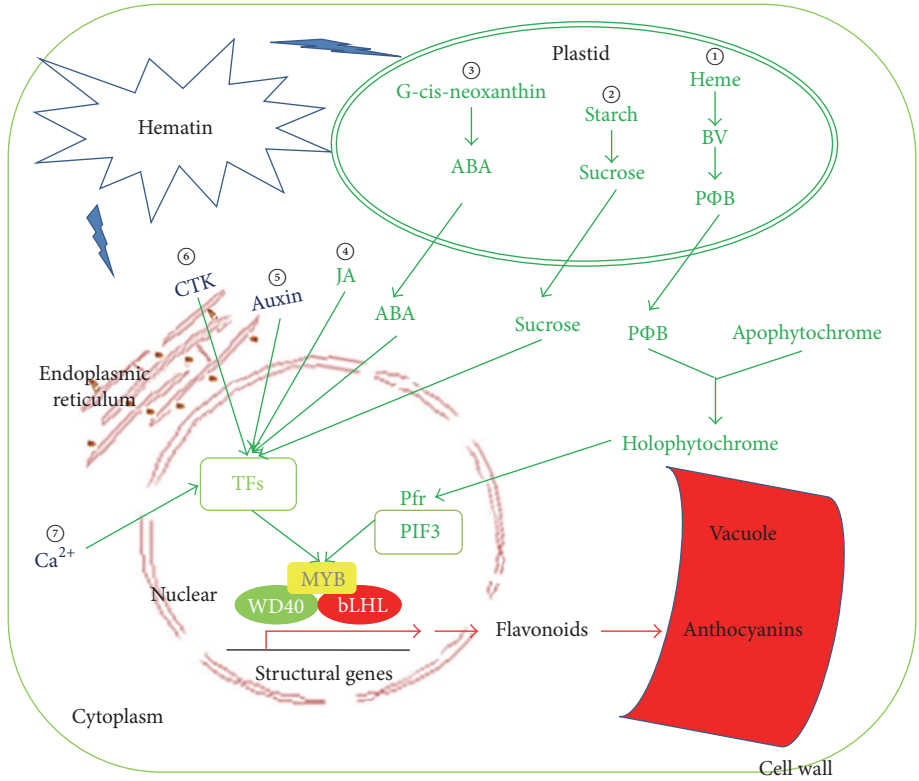


Figure 5: Genes and biological pathways that were involved in regulation of anthocyanin accumulation promoted by hematin. Outermost box represents the cell wall. Bilayer oval shape represents the plastid. The red-brown ring and network represent the nuclear and endoplasmic reticulum, respectively. The red crescent represents the vacuole. The box in the nucleus represents transcription factors (TFs). Yellow MYB, green WD40, and red bHLH represent a ternary transcription factor complex which transcribes anthocyanin biosynthesis genes. The circled numbers represent biological pathways involved in the regulation of anthocyanin accumulation promoted by hematin. ① represents the phytochrome regulation pathway. ② represents the starch and sucrose pathway. ③, ④, ⑤, and ⑥ represent phytohormones ABA, JA, Auxin, and CTK regulation pathways, respectively. ⑦ represents the Ca^{2+} regulation pathway. Pfr represents the far-red phytochrome. PIF3 represents phytochrome-interacting factor. Each anthocyanin biosynthesis regulatory pathway related gene is in Table 5.

in anthocyanin biosynthesis were significantly altered with the hematin treatment. This suggests that the physiological process of fruit color development is regulated through complex interactions among anthocyanin biosynthesis pathways, plant hormone signal transduction pathways, phytochrome signal transduction pathways, starch and sugar metabolic pathways, calcium pathways, and transcription factors (Figure 5). This study adds to the in-depth understanding of the fruit coloration process in strawberry.

Competing Interests

The authors declare that they have no conflict of interests.

Authors' Contributions

Jianwei Gao and Liangju Wang conceived and designed the research. Yi Li, Huayin Li, Fengde Wang, Jingjuan Li, and Yihui Zhang conducted experiments. Yi Li analyzed the data. Yi Li and Fengde Wang wrote the manuscript. All authors read and approved the manuscript. Yi Li and Huayin Li contributed equally to this work.

Acknowledgments

This work was supported by the Modern Agricultural Industrial Technology System Funding of Shandong Province, China (SDAIT-02-022-04), the China agriculture research system (CARS-25), and the National Science Foundation of China (31401820).

References

[1] K. Aaby, S. Mazur, A. Nes, and G. Skrede, "Phenolic compounds in strawberry (*Fragaria* x *ananassa* Duch.) fruits: composition in 27 cultivars and changes during ripening," *Food Chemistry*, vol. 132, no. 1, pp. 86–97, 2012.

[2] T. A. Colquhoun, L. A. Levin, H. R. Moskowitz, V. M. Whitaker, D. G. Clark, and K. M. Folta, "Framing the perfect strawberry: an exercise in consumer-assisted selection of fruit crops," *Journal of Berry Research*, vol. 2, no. 1, pp. 45–61, 2012.

[3] F. Giampieri, J. M. Alvarez-Suarez, L. Mazzoni et al., "The potential impact of strawberry on human health," *Natural Product Research*, vol. 27, no. 4, pp. 448–455, 2013.

[4] Y. S. Velioglu, G. Mazza, L. Gao, and B. D. Oomah, "Antioxidant activity and total phenolics in selected fruits, vegetables, and

grain products," *Journal of Agricultural & Food Chemistry*, vol. 46, no. 10, pp. 4113–4117, 1998.

[5] S. S. Hassellund, A. Flaa, S. E. Kjeldsen et al., "Effects of anthocyanins on cardiovascular risk factors and inflammation in pre-hypertensive men: a double-blind randomized placebo-controlled crossover study," *Journal of Human Hypertension*, vol. 27, no. 2, pp. 100–106, 2013.

[6] K. J. Meyers, C. B. Watkins, M. P. Pritts, and R. H. Liu, "Antioxidant and antiproliferative activities of strawberries," *Journal of Agricultural & Food Chemistry*, vol. 51, no. 23, pp. 6887–6892, 2003.

[7] S. Zafra-Stone, T. Yasmin, M. Bagchi, A. Chatterjee, J. A. Vinson, and D. Bagchi, "Berry anthocyanins as novel antioxidants in human health and disease prevention," *Molecular Nutrition & Food Research*, vol. 51, no. 6, pp. 675–683, 2007.

[8] L. Jaakola, "New insights into the regulation of anthocyanin biosynthesis in fruits," *Trends in Plant Science*, vol. 18, no. 9, pp. 477–483, 2013.

[9] K. Springob, J.-I. Nakajima, M. Yamazaki, and K. Saito, "Recent advances in the biosynthesis and accumulation of anthocyanins," *Natural Product Reports*, vol. 20, no. 3, pp. 288–303, 2003.

[10] A. Gonzalez, M. Zhao, J. M. Leavitt, and A. M. Lloyd, "Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings," *Plant Journal*, vol. 53, no. 5, pp. 814–827, 2008.

[11] K. Petroni and C. Tonelli, "Recent advances on the regulation of anthocyanin synthesis in reproductive organs," *Plant Science*, vol. 181, no. 3, pp. 219–229, 2011.

[12] H.-F. Jia, Y.-M. Chai, C.-L. Li et al., "Abscisic acid plays an important role in the regulation of strawberry fruit ripening," *Plant Physiology*, vol. 157, no. 1, pp. 188–199, 2011.

[13] C. Merchante, J. G. Vallarino, S. Osorio et al., "Ethylene is involved in strawberry fruit ripening in an organ-specific manner," *Journal of Experimental Botany*, vol. 64, no. 14, pp. 4421–4439, 2013.

[14] A. Maier, A. Schrader, L. Kokkelink et al., "Light and the E3 ubiquitin ligase COP1/SPA control the protein stability of the MYB transcription factors PAP1 and PAP2 involved in anthocyanin accumulation in *Arabidopsis*," *Plant Journal*, vol. 74, no. 4, pp. 638–651, 2013.

[15] H. Lange, W. Shropshire, and H. Mohr, "An analysis of phytochrome-mediated anthocyanin synthesis," *Plant Physiology*, vol. 47, no. 5, pp. 649–655, 1971.

[16] T. D. Elich and J. C. Lagarias, "Phytochrome chromophore biosynthesis: both 5-aminolevulinic acid and biliverdin overcome inhibition by gabaculine in etiolated *Avena sativa* L. seedlings," *Plant Physiology*, vol. 84, no. 2, pp. 304–310, 1987.

[17] Y. Han, W. Xuan, T. Yu et al., "Exogenous hematin alleviates mercury-induced oxidative damage in the roots of *Medicago sativa*," *Journal of Integrative Plant Biology*, vol. 49, no. 12, pp. 1703–1713, 2007.

[18] M. Koornneef, J. Cone, R. Dekens, E. O'Herne-Robers, C. Spruit, and R. Kendrick, "Photomorphogenic responses of long hypocotyl mutants of tomato," *Journal of Plant Physiology*, vol. 120, no. 2, pp. 153–165, 1985.

[19] M. J. Terry and R. E. Kendrick, "Feedback inhibition of chlorophyll synthesis in the phytochrome chromophore-deficient *aurea* and *yellow-green-2* mutants of tomato," *Plant Physiology*, vol. 119, no. 1, pp. 143–152, 1999.

[20] B. M. Parks and P. H. Quail, "Phytochrome-deficient *hy1* and *hy2* long hypocotyl mutants of *Arabidopsis* are defective in phytochrome chromophore biosynthesis," *Plant Cell*, vol. 3, no. 11, pp. 1177–1186, 1991.

[21] M. J. Terry, "Phytochrome chromophore-deficient mutants," *Plant, Cell & Environment*, vol. 20, no. 6, pp. 740–745, 1997.

[22] T. Muramoto, T. Kohchi, A. Yokota, I. Hwang, and H. M. Goodman, "The *Arabidopsis* photomorphogenic mutant *hy1* is deficient in phytochrome chromophore biosynthesis as a result of a mutation in a plastid heme oxygenase," *Plant Cell*, vol. 11, no. 3, pp. 335–347, 1999.

[23] M. D. Maines, "The heme oxygenase system: a regulator of second messenger gases," *Annual Review of Pharmacology and Toxicology*, vol. 37, no. 1, pp. 517–554, 1997.

[24] W. Xuan, F.-Y. Zhu, S. Xu et al., "The heme oxygenase/carbon monoxide system is involved in the auxin-induced cucumber adventitious rooting process," *Plant Physiology*, vol. 148, no. 2, pp. 881–893, 2008.

[25] R. Amooaghaie, F. Tabatabaei, and A.-M. Ahadi, "Role of hematin and sodium nitroprusside in regulating *Brassica nigra* seed germination under nanosilver and silver nitrate stresses," *Ecotoxicology and Environmental Safety*, vol. 113, pp. 259–270, 2015.

[26] Y. Liu, X. Li, L. Xu, and W. Shen, "De-etiolation of wheat seedling leaves: cross talk between heme oxygenase/carbon monoxide and nitric oxide," *PLoS ONE*, vol. 8, no. 12, Article ID e81470, 2013.

[27] T. Muramoto, N. Tsurui, M. J. Terry, A. Yokota, and T. Kohchi, "Expression and biochemical properties of a ferredoxin-dependent heme oxygenase required for phytochrome chromophore synthesis," *Plant Physiology*, vol. 130, no. 4, pp. 1958–1966, 2002.

[28] J.-W. Gao, J. Liu, B. Li, and Z. Li, "Isolation and purification of functional total RNA from blue-grained wheat endosperm tissues containing high levels of starches and flavonoids," *Plant Molecular Biology Reporter*, vol. 19, no. 2, pp. 185–186, 2001.

[29] R. Patro, S. M. Mount, and C. Kingsford, "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms," *Nature Biotechnology*, vol. 32, no. 5, pp. 462–464, 2014.

[30] Y.-X. Qi, Y.-B. Liu, and W.-H. Rong, "RNA-Seq and its applications: a new technology for transcriptomics," *Hereditas*, vol. 33, no. 11, pp. 1191–1202, 2011.

[31] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, no. 4, pp. 870–878, 2002.

[32] J. Ye, L. Fang, H. Zheng et al., "WEGO: a web tool for plotting GO annotations," *Nucleic Acids Research*, vol. 34, supplement 2, pp. W293–W297, 2006.

[33] H. Abdi, "The Bonferonni and Šidák corrections for multiple comparison," *Encyclopedia of Measurement and Statistics*, vol. 3, pp. 103–107, 2007.

[34] J. Li, Q. Ding, F. Wang, Y. Zhang, H. Li, and J. Gao, "Integrative analysis of mRNA and miRNA expression profiles of the tuberous root development at seedling stages in turnips," *PLoS ONE*, vol. 10, no. 9, Article ID e0137983, 2015.

[35] J. Bennett, D. Hondred, and J. C. Register, "Keeping qRT-PCR rigorous and biologically relevant," *Plant Cell Reports*, vol. 34, no. 1, pp. 1–3, 2015.

[36] F. Wang, L. Li, H. Li et al., "Transcriptome analysis of rosette and folding leaves in Chinese cabbage using high-throughput RNA sequencing," *Genomics*, vol. 99, no. 5, pp. 299–307, 2012.

[37] T. Shi, Z. Gao, L. Wang et al., "Identification of differentially-expressed genes associated with pistil abortion in Japanese apricot by genome-wide transcriptional analysis," *PLoS ONE*, vol. 7, no. 10, Article ID e47810, 2012.

[38] F. Wang, H. Li, Y. Zhang et al., "MicroRNA expression analysis of rosette and folding leaves in Chinese cabbage using high-throughput Solexa sequencing," *Gene*, vol. 532, no. 2, pp. 222–229, 2013.

[39] X. Gu, Y. Chen, Z. Gao, Y. Qiao, and X. Wang, "Transcription factors and anthocyanin genes related to low-temperature tolerance in rd29A: RdreB1BI transgenic strawberry," *Plant Physiology and Biochemistry*, vol. 89, pp. 31–43, 2015.

[40] J. G. Schaart, C. Dubos, I. R. De La Fuente et al., "Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (*Fragaria* × *ananassa*) fruits," *New Phytologist*, vol. 197, no. 2, pp. 454–467, 2013.

[41] L. Xie, Z. H. Wang, X. H. Cheng, J. J. Gao, Z. P. Zhang, and L. J. Wang, "5-Aminolevulinic acid promotes anthocyanin accumulation in *Fuji* apples," *Plant Growth Regulation*, vol. 69, no. 3, pp. 295–303, 2013.

[42] M. Daminato, F. Guzzo, and G. Casadoro, "A SHATTERPROOF-like gene controls ripening in non-climacteric strawberries, and auxin and abscisic acid antagonistically affect its expression," *Journal of Experimental Botany*, vol. 64, no. 12, pp. 3775–3786, 2013.

[43] X.-H. Ji, R. Zhang, N. Wang, L. Yang, and X.-S. Chen, "Transcriptome profiling reveals auxin suppressed anthocyanin biosynthesis in red-fleshed apple callus (*Malus sieversii* f. *niedzwetzkyana*)," *Plant Cell, Tissue and Organ Culture*, vol. 123, no. 2, pp. 389–404, 2015.

[44] J. Chen, L. Mao, W. Lu, T. Ying, and Z. Luo, "Transcriptome profiling of postharvest strawberry fruit in response to exogenous auxin and abscisic acid," *Planta*, vol. 243, no. 1, pp. 183–197, 2016.

[45] G. A. Martínez, A. R. Chaves, and M. C. Añón, "Effect of gibberellic acid on ripening of strawberry fruits (*Fragaria annanassa* Duch.)," *Journal of Plant Growth Regulation*, vol. 13, no. 2, pp. 87–91, 1994.

[46] P. K. Das, D. H. Shin, S.-B. Choi, S.-D. Yoo, G. Choi, and Y.-I. Park, "Cytokinins enhance sugar-induced anthocyanin biosynthesis in *Arabidopsis*," *Molecules & Cells*, vol. 34, no. 1, pp. 93–101, 2012.

[47] P. McAtee, S. Karim, R. Schaffer, and K. David, "A dynamic interplay between phytohormones is required for fruit development, maturation, and ripening," *Frontiers in Plant Science*, vol. 4, no. 3, article 79, 7 pages, 2013.

[48] E. Nambara and A. Marion-Poll, "Abscisic acid biosynthesis and catabolism," *Annual Review of Plant Biology*, vol. 56, pp. 165–185, 2005.

[49] Y. Jiang and D. C. Joyce, "ABA effects on ethylene production, PAL activity, anthocyanin and phenolic contents of strawberry fruit," *Plant Growth Regulation*, vol. 39, no. 2, pp. 171–174, 2003.

[50] S. Li, "Transcriptional control of flavonoid biosynthesis: fine-tuning of the MYB-bHLH-WD40 (MBW) complex," *Plant Signaling & Behavior*, vol. 9, Article ID e27522, 2014.

[51] D. R. Rudell, J. K. Fellman, and J. P. Mattheis, "Preharvest application of methyl jasmonate to 'Fuji' apples enhances red coloration and affects fruit size, splitting, and bitter pit incidence," *HortScience*, vol. 40, no. 6, pp. 1760–1762, 2005.

[52] A. Belhadj, N. Telef, C. Saigne et al., "Effect of methyl jasmonate in combination with carbohydrates on gene expression of PR proteins, stilbene and anthocyanin accumulation in grapevine cell cultures," *Plant Physiology and Biochemistry*, vol. 46, no. 4, pp. 493–499, 2008.

[53] E. Loreti, G. Povero, G. Novi, C. Solfanelli, A. Alpi, and P. Perata, "Gibberellins, jasmonate and abscisic acid modulate the sucrose-induced expression of anthocyanin biosynthetic genes in *Arabidopsis*," *New Phytologist*, vol. 179, no. 4, pp. 1004–1016, 2008.

[54] H. S. Seo, J. T. Song, J.-J. Cheong et al., "Jasmonic acid carboxyl methyltransferase: A key enzyme for jasmonate-regulated plant responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4788–4793, 2001.

[55] M. T. McDowell and J. C. Lagarias, "Purification and biochemical properties of phytochromobilin synthase from etiolated oat seedlings," *Plant Physiology*, vol. 126, no. 4, pp. 1546–1554, 2001.

[56] G.-Q. Fu, S. Xu, Y.-J. Xie et al., "Molecular cloning, characterization, and expression of an alfalfa (*Medicago sativa* L.) *heme oxygenase-1* gene, *MsHO1*, which is pro-oxidants-regulated," *Plant Physiology and Biochemistry*, vol. 49, no. 7, pp. 792–799, 2011.

[57] X.-Y. Chen, X. Ding, S. Xu et al., "Endogenous hydrogen peroxide plays a positive role in the upregulation of heme oxygenase and acclimation to oxidative stress in wheat seedling leaves," *Journal of Integrative Plant Biology*, vol. 51, no. 10, pp. 951–960, 2009.

[58] J. Shin, E. Park, and G. Choi, "PIF3 regulates anthocyanin biosynthesis in an HY5-dependent manner with both factors directly binding anthocyanin biosynthetic gene promoters in *Arabidopsis*," *Plant Journal*, vol. 49, no. 6, pp. 981–994, 2007.

[59] S. Wang, D. Pan, X. Lv et al., "Proteomic approach reveals that starch degradation contributes to anthocyanin accumulation in tuberous root of purple sweet potato," *Journal of Proteomics*, vol. 143, pp. 298–305, 2016.

[60] H.-F. Jia, C.-L. Li, Y.-M. Chai, Y. Xing, and Y. Shen, "Sucrose promotes strawberry fruit ripening by stimulation of abscisic acid biosynthesis," *Pakistan Journal of Botany*, vol. 45, no. 1, pp. 169–176, 2013.

[61] W. Xu, H. Peng, T. Yang et al., "Effect of calcium on strawberry fruit flavonoid pathway gene expression and anthocyanin accumulation," *Plant Physiology & Biochemistry*, vol. 82, pp. 289–298, 2014.

[62] Z.-W.-S. Wan and Z. Singh, "Exogenous application of prohexadione-calcium promotes fruit colour development of '*Cripps Pink*' apple," *Acta Horticulturae*, vol. 1012, pp. 219–225, 2012.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2016, Article ID 2895303, 15 pages
http://dx.doi.org/10.1155/2016/2895303

Research Article

Differential Gene Expression during Larval Metamorphic Development in the Pearl Oyster, *Pinctada fucata*, Based on Transcriptome Analysis

Haimei Li,[1,2] Bo Zhang,[1] Guiju Huang,[1] Baosuo Liu,[1] Sigang Fan,[1] Dongling Zhang,[3] and Dahui Yu[1]

[1]*Key Laboratory of South China Sea Fishery Resources Exploitation & Utilization of Ministry of Agriculture, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, Guangdong 510300, China*
[2]*Shanghai Ocean University, Shanghai 201306, China*
[3]*Jimei University, Xiamen, Fujian 361021, China*

Correspondence should be addressed to Dahui Yu; pearlydh@163.com

Received 28 June 2016; Revised 26 August 2016; Accepted 20 September 2016

Academic Editor: Quanhu Sheng

P. fucata experiences a series of transformations in appearance, from swimming larvae to sessile juveniles, during which significant changes in gene expression likely occur. Thus, *P. fucata* could be an ideal model in which to study the molecular mechanisms of larval metamorphosis during development in invertebrates. To study the molecular driving force behind metamorphic development in larvae of *P. fucata*, transcriptomes of five larval stages (trochophore, D-shape, umbonal, eyespots, and spats) were sequenced using an Illumina HiSeq™ 2000 system and assembled and characterized with the transcripts of six tissues. As a result, a total of 174,126 unique transcripts were assembled and 60,999 were annotated. The number of unigenes varied among the five larval stages. Expression profiles were distinctly different between trochophore, D-shape, umbonal, eyespots, and spats larvae. As a result, 29 expression trends were sorted, of which eight were significant. Among others, 80 development-related, differentially expressed unigenes (DEGs) were identified, of which the majority were homeobox-containing genes. Most DEGs occurred among trochophore, D-shaped, and UES (umbonal, eyespots, and spats) larvae as verified by qPCR. Principal component analysis (PCA) also revealed significant differences in expression among trochophore, D-shaped, and UES larvae with ten transcripts identified but no matching annotations.

1. Introduction

Metamorphosis is a series of key steps in the process of larval development, the success of which affect the survival of the organism. Metamorphosis is prevalent in insects, amphibians, some fishes, and many marine invertebrates, such as barnacles, sponges, shellfishes, shrimps, and echinoderms. Similar to most benthic marine invertebrates, the pearl oyster (*Pinctada fucata*) has a microscopic, free-swimming larval phase in their complex life cycle [1]. Oyster larvae spend several weeks in the water column before attaining competency to attach and metamorphose, commencing their sessile life. The developmental processes of *P. fucata*, from swimming larvae to sessile spats, have been classified into six stages: fertilized egg, trochophore, D-shaped, umbonal, juvenile, and adult stages [2]. In oysters, the transition from free-swimming larvae to the attached juvenile form often requires morphological, physiological, structural, and functional changes, which are under genetic regulatory control [3]. Therefore, the identification of key developmental genes involved in the metamorphosis of *P. fucata* larvae, as well as characterizing their expression patterns, is important to understand the molecular mechanism of metamorphic development of this economically important species.

Numerous studies have been conducted to explore the mechanisms of hormones, neurotransmitters, genes, and signaling pathways that regulate larval metamorphic development. Some studies have demonstrated that eight

superfamily genes showed differential expression during the metamorphosis of *Ciona intestinalis* [4–7]. Several homeobox-containing genes were found to be responsible for larval metamorphic development in *Haliotis rufescens* [8–10]. In addition, abnormal dopamine and adrenaline were observed in the larval attaching stage of the Pacific oyster, *Crassostrea gigas* [11], while a different study observed increased expression of a molluscan growth and differentiation factor (mGDF) in the metamorphosing stage of the same organism [12]. These findings indicate the diversity of genes involved in the transitions of larval forms.

However, previous studies have focused on changes in a small number of genes and have provided a fragmented view of the genetic modulation of larval metamorphosis. Recent developments in sequencing technology have allowed for the development of new genomic tools, which can provide a more global view of changes in gene expression over the course of larval developmental stages [13–16]. In terms of genome-wide studies, transcriptome analyses are considered to be an ideal choice for obtaining comprehensive information regarding animal development and growth [17, 18]. For *P. fucata*, the draft genome [19] and tissue transcriptomes [20–23] have been recently reported. Based on the transcriptomic sequences from a mixture of nine developmental stages of *P. fucata* [19], biomineralization-related gene expression profiles during larval development have been investigated [24, 25] and genes involved in body patterning [26], transcription factors [27], and homeobox genes [28] have been identified. Nonetheless, developmentally important genes and their expression patterns during the larval stages of developing *P. fucata* have not been systematically studied at the transcript level to date. In the present study, the transcriptomes of five larval stages (trochophore, D-shape, umbonal, eyespots, and spats) and six tissues (gill, adductor muscle, hepatopancreas, mantle, hemocytes, and pearl sac) from *P. fucata* were sequenced using Illumina HiSeq 2000, with an emphasis on the molecular mechanisms underlying larval metamorphic development. This study aims to provide a valuable insight into the mechanisms of genetic modulation over the course of larval metamorphic development for *P. fucata* as well as for other molluscan species.

2. Materials and Methods

2.1. Larval Culture and Sample Collection. Larvae of *P. fucata* were bred (using several females and males of a selectively bred F3 generation as parents) through artificial insemination on March 10, 2013, in Sanya, Hainan Island, China, as described by Fujimura et al. [2]. Fertilized eggs were incubated in a 1000 L tank at 24°C. After removing nondeveloping embryos and dead larvae, trochophore, D-shaped, umbonal, eyespots, and spats larvae stages were harvested with filtering net at 12 h, 36 h, 11.5 d, 18.5 d, and 23.5 d after fertilization, respectively, and immediately preserved in RNA later (TaKaRa Bio Inc) until RNA extraction. Meanwhile, RNAs of six tissues (gill, adductor muscle, hepatopancreas, mantle, hemocytes, and pearl sac) of three other adult animals were sequenced for a more robust assembly.

2.2. RNA Extraction and cDNA Library Preparation. Following the manufacturer's instructions, total RNA was extracted from five developmental stages (each stage with thousands of larvae) and six tissues using Trizol and RNAs of each type of tissues of the three individuals were mixed by equal weight. RNA integrity and quantity were confirmed by lab-on-chip analysis using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and visualized on a 1% agarose gel. Then cDNA was synthesized using the mRNA fragments as templates as usual and was sequenced by the BGI (Shenzhen, China) using the Illumina HiSeq 2000 system (San Diego, CA, USA) (PE100).

2.3. Sequence Assembly and Annotation. After filtering out low quality sequences (containing more than 5% ambiguous "N" nucleotides or >20% $Q \leqq 10$ reads) and the removal of adapters from raw data, clean sequence data was assembled into unigenes using Trinity software and subsequently clustered by TGICL v2.1 (-l 40 -c 10 -v 20) [29]. Phrap (-repeat_stringency 0.95 -minmatch 35 -minscore 35) (Release 23.0) was used to produce the longest sequence possible (http://www.phrap.org/). Assembled unigenes were annotated based on the Nr, Swissprot, KEGG, and COG databases. The sequence direction and amino sequence of the predicted coding region (CDS) of unannotated unigenes were determined using ESTScan with default settings [30]. Functional annotations and classifications were performed by using Blast2GO [31] and WEGO [32] (E value threshold 1×10^{-5}), respectively.

2.4. Normalization and Quantification of Gene Expression. Sequencing reads were mapped to the assembled reference sequence using SOAP aligner/soap2 (-m 0 -x 500 -s 40 -l 35 -v 5 -r 2) [33], a tool designed specifically to assemble short sequence alignments. The coverage of reads from a given gene was used to calculate the expression level of that gene, which was measured by fragments per kilobase exon per million fragments (FPKM) [34], with the following formula:

$$\text{FPKM} = \frac{10^6 C}{NL/10^3}, \tag{1}$$

where FPKM is the expression level of a unigene, C is the number of fragments that uniquely aligned to the unigene, N is the total number of fragments that uniquely aligned to all unigenes, and L is the number of bases in the CDS of the unigene. The FPKM method eliminates the influence of sequences of differing lengths and coverage level on the calculation of gene expression. Therefore, the calculated gene expression can be directly used for comparing the difference in gene expression between samples.

2.5. Differential Gene Expression (DEGs) across Developmental Stages. Differential gene expression among different larval stages was carried out via principal component analysis (PCA) using the R package (http://www.r-project.org/) according to the manual. The pairwise differential expression conducted by edgeR, with a threshold of the false discovery rate (FDR) ≤ 0.001 and an absolute value of $\log 2$ Ratio ≥ 1, was used to judge the significance of differences in gene expression. Trends in the expression of all differentially

Table 1: Primers for genes used for qPCR verification.

Trend	Unigene code	Annotated gene	Primer
0	Unigene27615_All	*Brachyury*	qBran-S: 5′ GCCAAAGAAAGACCAGAAGG 3′ qBran-A: 5′ TCAGGCTAAGGCGATCACAA 3′
0	Unigene27517_All	*Six3*	qSix3-S: 5′ ATACAGGGTGAGGAAGAAGT 3′ qSix3-A: 5′ TTATCTCCGCCTTGCTGTTG 3′
3	Unigene23318_All	*Engrailed*	qEng-S: 5′ TAGACAGAGCATCGCCTTTA 3′ qEng-A: 5′ TTGTGATTTAACTGCCTGCT 3′
3	Unigene41009_All	*Pax-7*	qPax-S: 5′ GCGGAAACAGATGGGAAGCA 3′ qPax-A: 5′ ACCGAATGACGGAAACGACT 3′
26	CL664.Crontig4_All	*MAPK*	qMAPK-S: 5′ TTTACTCCAAACAGCCCTAC 3′ qMAPK-A: 5′ TTGCTATCTGGTCCACTTCA 3′
29	CL7953.Contig2_All	*Notch*	qNotch-S: 5′ CCAGCCACGGTATCCAAGTA 3′ qNotch-A: 5′AGCCTCGAACAGAATATCCACT 3′
29	CL1306.Contig2_All	*Wnt 1*	qWnt1-S: 5′ TGATGCCTACGGTAAATACG 3′ qWnt1-A: 5′ TAACCTTGAGGTGGGAGAAC 3′
29	Unigene27337_All	*Lox2*	qLox2-S: 5′ CTACCCGAGTTGAATGTGGG 3′ qLox2-A: 5′ GAAAGTAAGACGGACGAGCC 3′

expressed genes were sorted using STEM (Short Time-Series Expression Miner, v1.3.8) [35]. Functional annotation and classification of genes involved in significant trends were performed by using Blast2GO [31] and WEGO [32], respectively. The enriched metabolic pathways or signal transduction pathways of genes were identified based on the KEGG database [36].

2.6. Identification and Expression Profile of Genes Involved in the Larval Metamorphic Development of P. fucata. According to annotations by Nr and Swissprot, development-related genes were identified with those that had been previously identified as keywords in the significant trends from the prior step. If several unigenes were assigned to the same reference gene, the sequence with the lowest *E* value (Nr and Swissprot annotation *E* value) was selected as a representative. Then, the heatmap 2 module of the gplots package in R (https://cran.r-project.org/web/packages/gplots/index.html) was used to perform the clustering analysis of gene expression on the normalized, filtered sequences to identify genes that were significantly different among the five developmental stages.

2.7. qPCR Verification of Expression Trends of Development-Related Genes. In order to verify the integrity of the transcriptome sequences and the expression levels as revealed by RNA-Seq, eight development-related genes were selected randomly for qPCR verification. The genes and respective primers are given in Table 1. qPCR was performed using an Eppendorf real-time- (RT-) PCR system (Eppendorf, Hamburg, Germany) using a SYBR(R) Premix Ex TaqTM kit (TaKaRa) according to the manufacturer's protocol. Transcript levels of target genes were normalized against the level of a reference gene (18S rRNA). The qRT-PCR reactions were performed under the following conditions: 94°C for 5 min (one cycle), 94°C for 20 s, 50°C to 60°C for 20 s, and 72°C for 20 s (50 cycles). The comparative CT method ($2^{-\Delta\Delta CT}$ method) was used to determine the relative mRNA abundance [37].

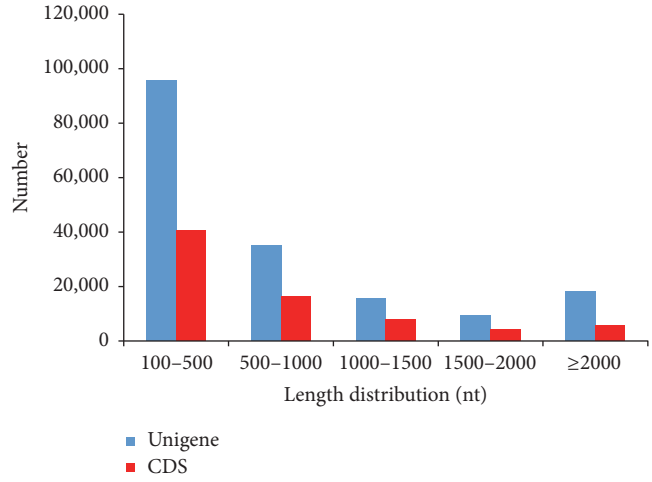


Figure 1: Length distribution of unigenes and CDS.

3. Results

3.1. Sequence Assembly and Annotation. Over 55 million reads per sample were generated with a base call accuracy (Q20) of over 97%. The number of contigs varied from 118,010 to 215,808, with a median length (N50) of 352 to 582 bp (Table 2). The number of unigenes varied from 51,102 to 113,516 among samples, with the mean length ranging from 536 to 689 bp, while N50 ranged from 495 to 1,025, respectively. In total, 174,126 unigenes were assembled with a mean length of 866 bp and an N50 of 1, 569 (based on 11 samples). Most unigenes were 100–500 bp long, and 26% were greater than 1,000 bp (Figure 1).

Table 2: Summary statistics of sequence assembly from 11 samples of the pearl oyster, *Pinctada fucata*.

Sample	Total raw reads	Total clean reads	Q20	Number of contigs	ContigN50	Number of unigenes	Mean length	N50
Gill	59,492,360	53,662,442	97.67%	215,808	380	113,516	592	977
Adductor muscle	58,602,062	53,889,264	97.42%	127,634	347	76,240	415	495
Hepatopancreas	57,854,582	52,672,774	97.75%	163,089	398	85,839	544	811
Pearl sac	56,926,624	52,155,950	97.49%	149,236	582	97,501	545	827
Mantle	58,610,258	52,433,102	97.72%	183,633	384	100,679	567	900
Hemocytes	54,707,500	51,751,784	98.54%	178,460	509	96,469	599	1025
Trochophore	59,887,806	54,413,910	97.96%	175,174	399	75,400	584	785
D-shaped	58,511,662	51,746,334	97.98%	190,135	485	88,830	626	886
Umbonal	66,858,916	55,046,652	97.88%	118,010	352	51,102	536	683
Eyespots	58,534,394	52,943,288	97.88%	182,671	504	84,045	649	937
Spats	60,561,378	54,999,258	97.85%	180,265	521	82,133	689	1022
All						174,126	866	1569

Table 3: Summarized statistics of the functional annotation in 11 samples of the pearl oyster, *Pinctada fucata*.

Database	Number of annotated genes
Nr annotation	38857
Swissprot annotation	49580
Annotation unigenes for KEGG	43753
Annotation unigenes for GO	13465
Annotation unigenes for COG	23754
CDS	60,946
CDS by ESTScan	13,966
Total	60999

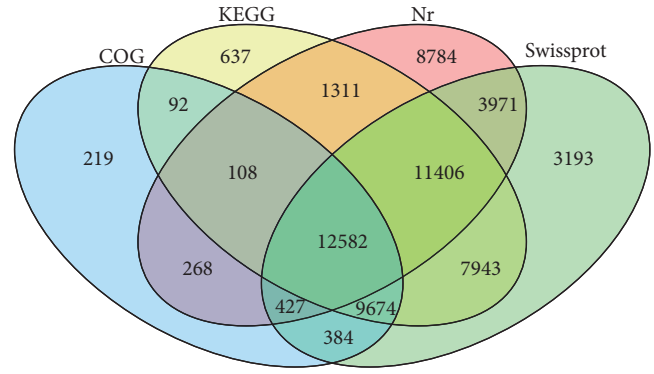


Figure 2: Number of genes annotated by COG, KEGG, Nr, and Swissprot terms based on five larval stages and six tissues of *Pinctada fucata*.

In total, 60,999 unigenes were annotated (Table 3) and 74,912 CDSs (43.02%) were predicted (13,966 predicted by ESTScan) (Figure 1, Table 3). Different databases annotated different numbers of unigenes (Table 3), where the most unigenes (49,580) were annotated by Swissprot database and the least (13,465) by the GO database (Table 3). Numbers of specific and shared unigenes annotated by COG, KEGG, Nr, and Swissprot terms can be visualized in Figure 2. Among them, 12,582 unigenes were annotated by the four databases and 8,784 were annotated specifically by Nr (Figure 2). Both KEGG and Swissprot analyses shared the most unigenes (41,605) and COG and Nr shared the least (13,385).

The 23,754 COG-annotated unigenes can be further classified into 25 functional groups, half of which were sorted into the "general function" group (Figure 3). The GO analysis revealed that 10,165 unigenes were attributed to biological process, 8,442 unigenes to cell components, and 10,588 unigenes to molecular function (Figure 4). The top 26 KEGG pathways are summarized in Table 4. Most unigenes (5,184 out of 43,753) were involved in metabolic pathways and 1,401 unigenes were involved in calcium signaling pathways, some of which may be involved in shell formation. Finally, many unigenes in the top 26 KEGG pathways were involved in immune pathways.

3.2. Differential Gene Expression (DEGs) and Expression Trends during Developmental Stages. Principal component analysis (PCA) revealed that differences in the expression of unigenes were vast among trochophore, D-shaped, and UES (umbonal, eyespots, and spats) larvae, but small within UES stages (Figure 5(a)). Based on gene effects, measured by the first principal component value, a total of 10 transcripts with unknown functions were identified to be key factors involved in the larval development of *P. fucata*. Differences in the gene expression of these transcripts were the greatest in trochophore, D-shaped, and UES larvae. They were relatively highly expressed in trochophore larvae and then downregulated during the D-shaped stage, and some were subsequently upregulated during the UES stage, including Unigens23340_All, Unigene8217_All, Unigene50061_All, and Cl616_All (Figure 5(b)).

The numbers of up- and downregulated unigenes were also much greater during early stage transitions (Figure 6), consistent with the results of the PCA. From trochophore to D-shape larvae, there were 18,725 unigenes upregulated and 13,162 downregulated. In total, there were 57,228 DEGs among the five developmental stages (Figure 7). Additionally, 17,609 genes were preferentially expressed at a single developmental stage, which indicates that they play an important role in the corresponding developmental stage, while 39,619 were expressed preferentially during more than two stages.

Table 4: Top 26 KEGG pathways.

Pathway	Count (43,753)	Pathway ID
Metabolic pathways	5184	ko01100
Regulation of actin cytoskeleton	2302	ko04810
Vascular smooth muscle contraction	2176	ko04270
Focal adhesion	2142	ko04510
Pathways in cancer	1607	ko05200
Tight junction	1544	ko04530
Hypertrophic cardiomyopathy (HCM)	1426	ko05410
Dilated cardiomyopathy	1402	ko05414
Calcium signaling pathway	1401	ko04020
Amoebiasis	1384	ko05146
Tuberculosis	1379	ko05152
RNA transport	1336	ko03013
Salmonella infection	1333	ko05132
Neuroactive ligand-receptor interaction	1304	ko04080
Epstein-Barr virus infection	1240	ko05169
Phagosome	1239	ko04145
Spliceosome	1231	ko03040
Purine metabolism	1191	ko00230
Vibrio cholera infection	1157	ko05110
Endocytosis	1139	ko04144
Huntington's disease	1121	ko05016
Viral myocarditis	1090	ko05416
MAPK signaling pathway	1046	ko04010
Cardiac muscle contraction	1044	ko04260
Gastric acid secretion	1019	ko04971
Ubiquitin mediated proteolysis	1014	ko04120

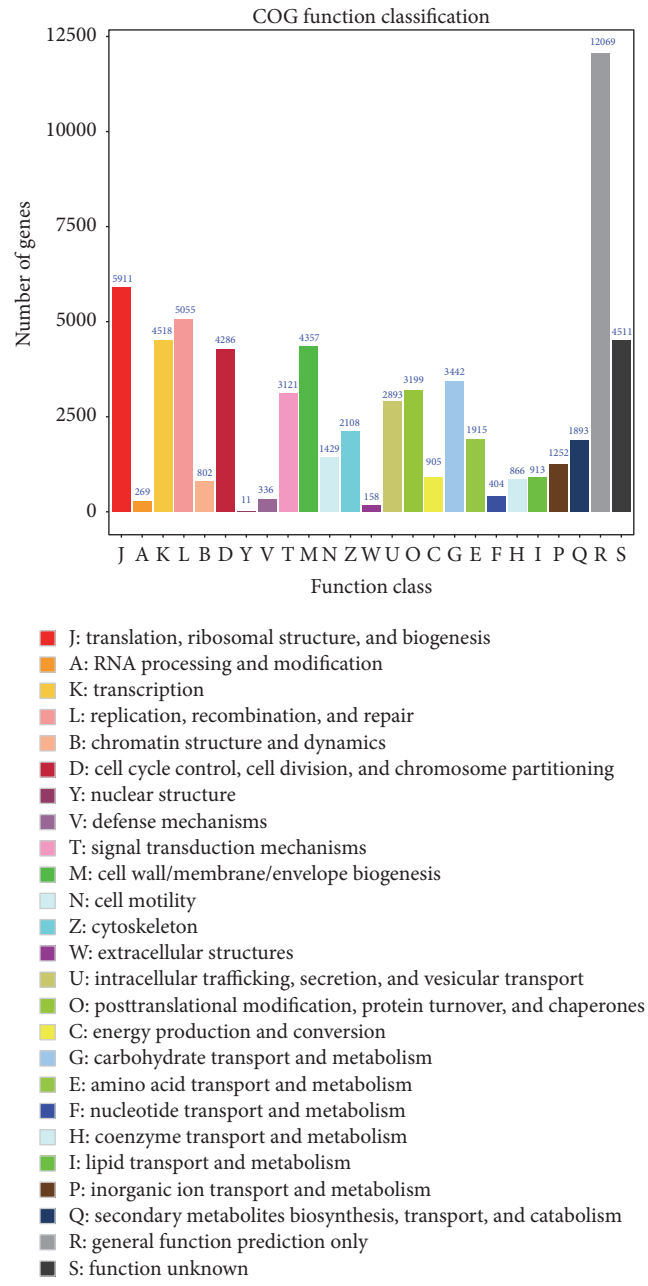


Figure 3: COG function classification of all unigenes.

A total of 20,518 genes were differentially expressed in all five of the development stages. All differentially expressed genes were sorted into 29 expression trends (Figure 8), of which eight trends were significant, comprising over 45% of the total DEGs. Furthermore, 6,653 unigenes were expressed highly only during the trochophore stage. Across the five stages, 3,340 unigenes were expressed in an increasing pattern, while 2,631 unigenes were expressed in a decreasing pattern.

3.3. Functional Enrichment Analysis. A functional enrichment analysis of the unigenes from the eight significant trends showed that there were 104, 54, and 46 GO terms for biological processes, molecular functions, and cellular components, respectively, identified for GO function enrichment (see Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/2895303), and 272 pathways identified for KEGG pathway enrichment (Supplementary Table 2). For GO enrichment data, trends 0, 2, 3, 24, 26, 28, and 29 were involved in biological processes, where most unigenes belonged to trend 3 and were related to various metabolic processes. Trends 0, 3, 24, 28, and 29 were involved in molecular function, where most unigenes fell within trends 0, 3, and 28, and were related to binding and catalytic activity. Only trends 0, 3, 28, and 29 were involved in cellular components, and most unigenes fell within trend 3 and were involved in processes related to membranes and organelles.

Trends 0, 2, 3, 24, 26, 27, 28, and 29 were implicated in KEGG pathway enrichment. For 272 significant enriched pathways, 81 pathways were observed in trend 29, 77 in trend 28, 30 in trend 27, and 27 in trend 26 (Supplementary Table 2). In trends 28 and 29, most unigenes were involved in immune responses.

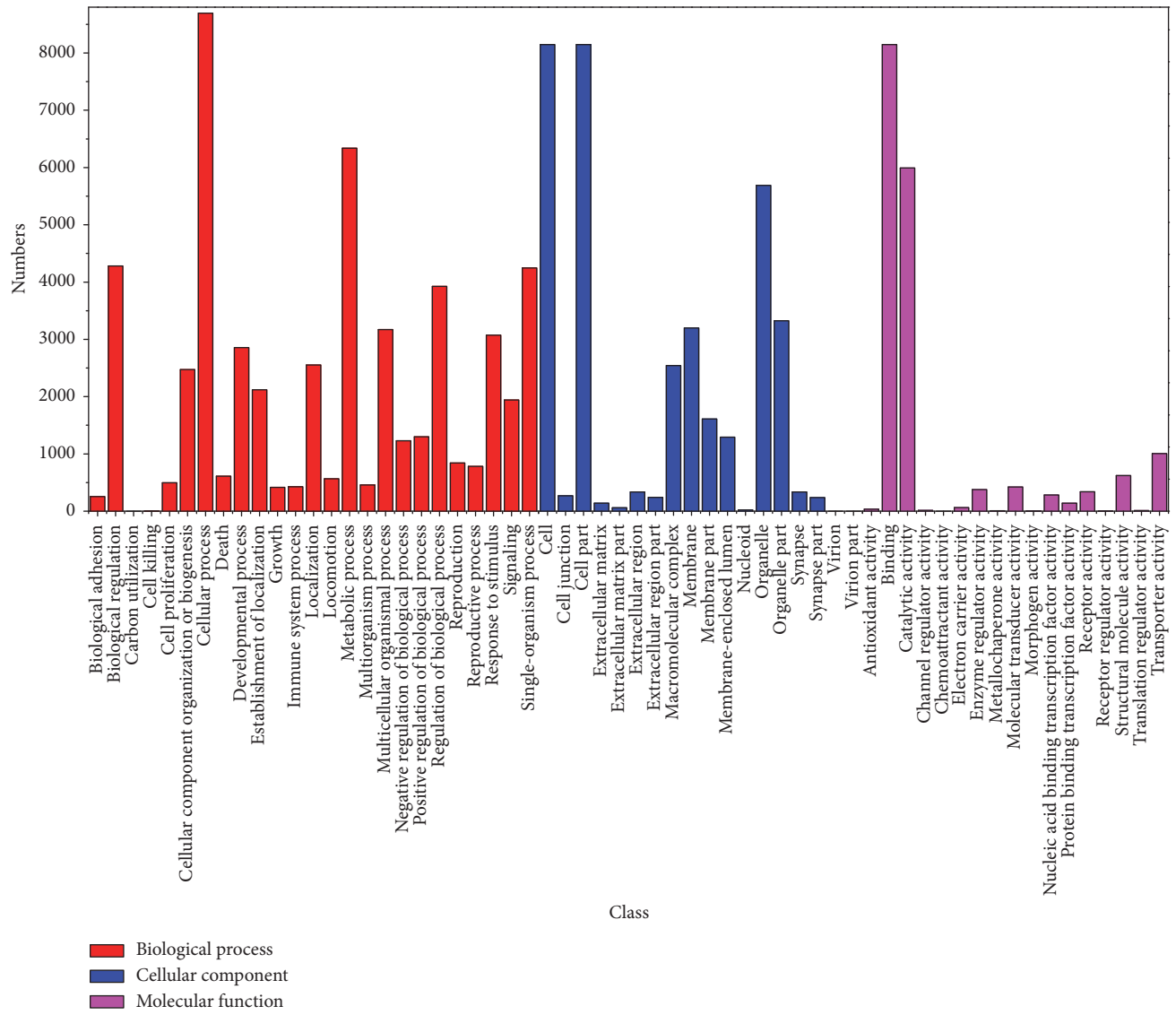


Figure 4: GO categories of unigenes. 13,465 of 174,126 unigenes were assigned to GO annotation and divided into three categories: biological processes, cellular components, and molecular functions.

In trend 0, GO enrichment showed that macromolecule metabolic processes were the dominant groups in biological process, followed by positive regulation of biological process. For pathways involved in molecular function, DNA binding was the most representative category, while for cellular component pathways, all of the genes participate in processes integral to the inner mitochondrial membrane and are intrinsic to the inner mitochondrial membrane. In the KEGG category, spliceosomes were prevalent, followed by genes involved in the cell cycle.

In trend 3, GO enrichment data showed that 6,653 DEG unigenes were further categorized into 43 functional groups; among them, macromolecule metabolic processes were the dominant groups in biological process, followed by cellular macromolecule metabolic processes. In the molecular function category, a high percentage of genes came from the binding and protein binding groups. Spliceosomes were the most representative, followed by RNA transport and regulation of the actin cytoskeleton.

In trends 27 and 28, GO enrichment reveled that there were no significant categories. However, the calcium signaling pathway, hedgehog signaling pathway, and insulin signaling pathway were significantly enriched in the KEGG database, as they are all involved in early development. In trend 28, small molecule metabolic processes were the dominant group in biological process followed by ion transport. Catalytic activity was the most prevalent in the molecular function category, followed by transporter activity and transmembrane transporter activity. In cellular component pathways, membrane was the most representative, followed by plasma membrane. In KEGG enrichment categories, we also found genes related to the calcium signaling pathway in trend 27.

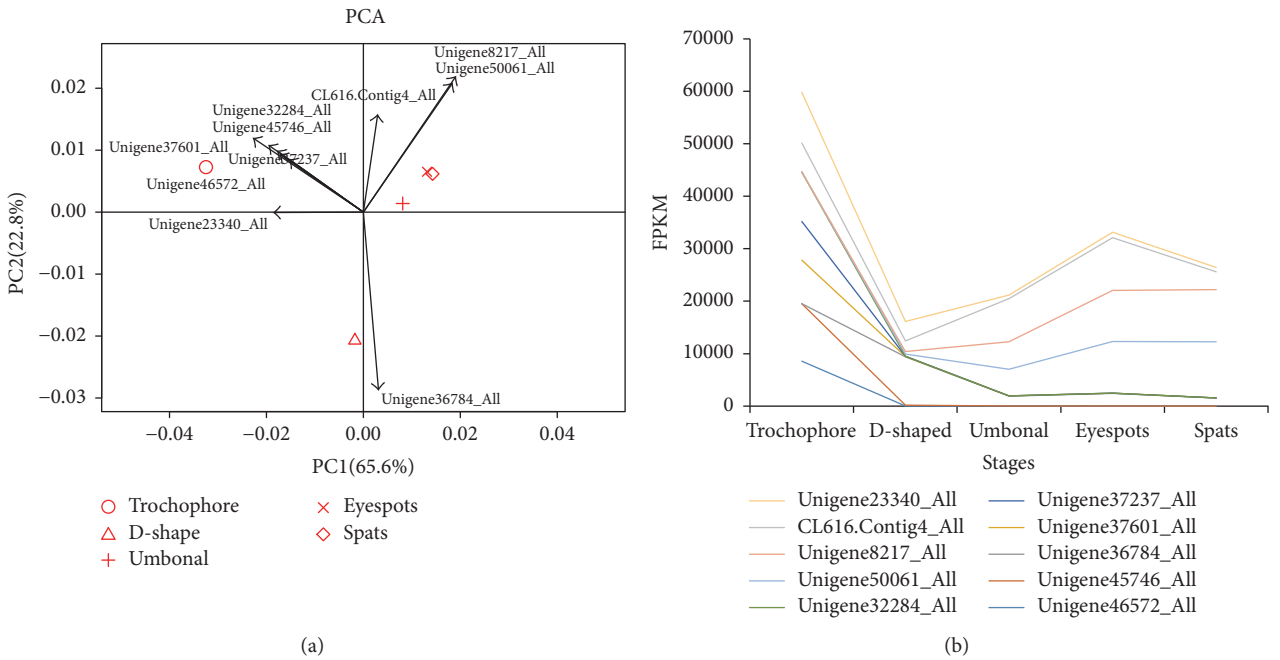


Figure 5: The distinction between five developmental stages as indicated by (a) principal component analysis and (b) expression levels of 10 representative genes, identified to be responsible for the distinction among the developmental stages.

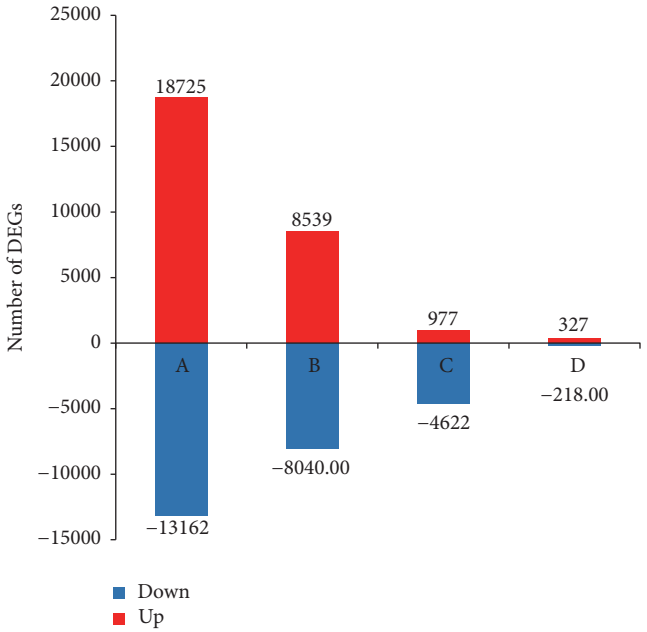


Figure 6: Numbers of up- (red) and downregulated (blue) unigenes. The numbers on column indicate the quantity of up- (red) and downregulated (blue) genes. The results of four comparisons are shown. The signal intensities of each feature of the DEGs are plotted on a logarithmic scale. Statistical criteria for designation of genes as up- or downregulated are outlined in the methods.

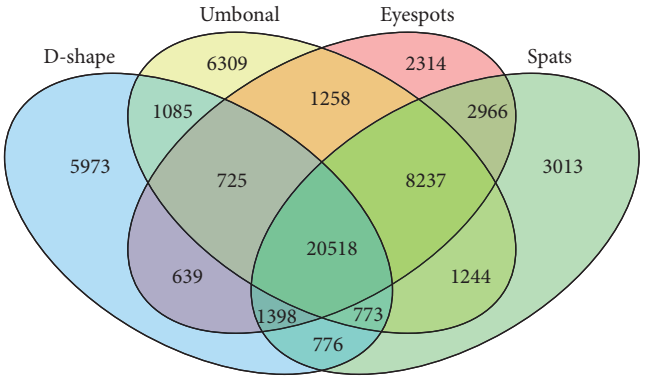


Figure 7: Specific and shared genes in five developmental stages of the pearl oyster, *Pinctada fucata*.

In trend 29, translational elongation was the only enriched category for biological processes, while three categories were enriched in molecular function, including genes involved in oxidoreductase activity, catalytic activity, and lyase activity. In cellular component pathways, five categories were enriched, while vacuole was the dominant group, followed by lytic vacuole and lysosome groups. Genes in trend 29 were enriched in only one KEGG pathway, translational elongation, with a significant *E* value.

3.4. Identification and Expression Profiling of Genes Involved in the Larval Metamorphic Development of P. fucata. In total, 80 development-related candidate DEGs were identified and summarized in Table 5, which can be mapped to known developmentally important genes, including several homeobox genes, and can be sorted into 10 trends: trend 0 (25 unigenes), trend 28 (16), trend 3 (15), trend 29 (9), and six other trends (1–5). Cluster analyses suggested that most development-related candidate genes were highly expressed

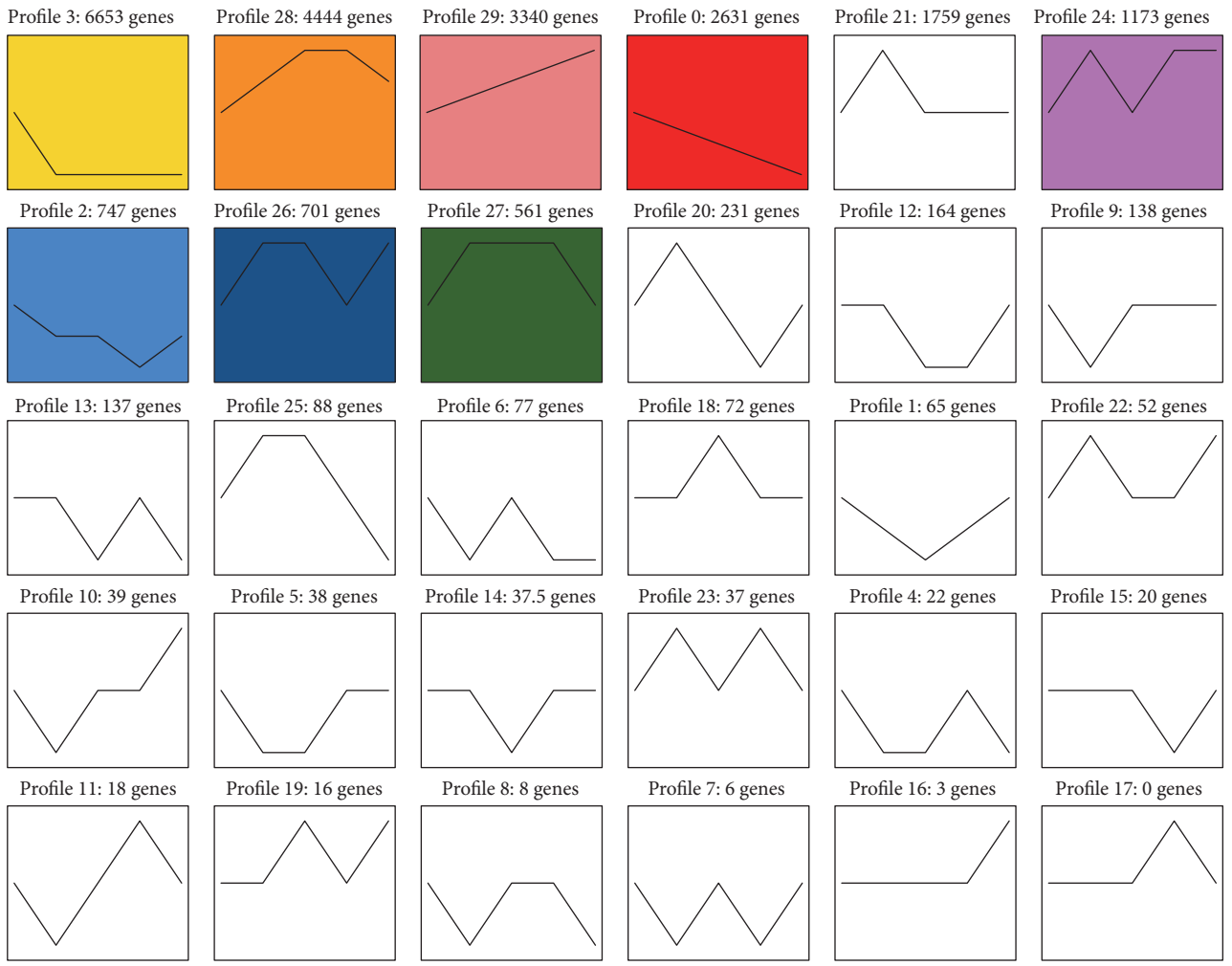


Figure 8: Expression trends of unigenes across trochophore, D-shaped, umbonal, eyespots, and spats larval stages of *Pinctada fucata*. The profiles were ordered based on the P value of the number (at bottom-left corner) of genes assigned versus expected. Color square frames denote significant profiles ($P \leq 0.01$). Each graph displays the mean pattern of expression (black lines) of the profiled genes. The number of profiles in each cluster is indicated in the top left corner. The x-axis represents stages and the y-axis represents log 2-fold change of gene expression.

in the early developmental stages (Figure 9), including *engrailed-2-B*, *pax* family, *fox* family members *e1* and *p1*, *Wnt-4,* and *BMP3/3B* upregulated in the trochophore stage and *LIM*, *foxg1*, *Hox3*, *bicaudal*, *hedgehog*, *EGFR*, *foxl2*, and *bmp 2b* genes upregulated from the D-shaped stage until the eyespots stage. In the spats stage, *wnt1* and *notch-like protein 2* gene were upregulated. The qPCR showed that the trends in the expression of selected genes (Figure 10) were consistent with the expression trends indicated by the trend analysis of RNA-Seq data (Figure 9), indicating that the sequence data in our study are reliable.

4. Discussion

Not only does the pearl oyster, *P. fucata*, make an ideal model organism for studies of biomineralization, but also it is a good model to study the early stage metamorphic development of invertebrates. In this study, we sequenced the transcriptomes of five developmental stages in *P. fucata*, with the aim of developing a better understanding of the molecular mechanisms driving the change of one larval stage to the next during early life history. In our study, the *de novo* assembly was performed with six tissue transcriptomes, as the draft genome of *P. fucata* is not complete [19]. As a result, we obtained 174,126 unigenes, with a mean length of 866 bp. A total of 60,999 unigenes (35%) were annotated, a value slightly higher than previous reports [21–23, 38]. Poor annotation efficiencies have been widely prevalent in many marine organisms, likely owing limited genomic resources from aquaculture species in public databases to date [21–23, 38]. Alternatively, poor annotation efficiencies could be the result of the short length of the assembled unigene sequences [22] and great divergence among the genomes of marine organism. Similar scenarios have been reported in other marine organisms [39, 40]. In the KEGG annotation, we observed that many pathways were

Table 5: Early development-related DEGs and their expression trends in *Pinctada fucata* (80).

Unigene ID	Annotated gene	Reference species	*E* value	Trend
Unigene41154_All	Nanos-like protein 1	*Crassostrea gigas*	$5.00E-73$	0
Unigene40485_All	EGF-like 4	*Crassostrea gigas*	$9.00E-41$	0
Unigene27615_All	brachyury	*Saccostrea kegaki*	0	0
Unigene40515_All	TBX6	*Crassostrea gigas*	$2.00E-49$	0
CL19363.Contig1_All	foxn1	*Homo sapiens*	$2.00E-48$	0
Unigene36457_All	foxb1	*Crassostrea gigas*	$6.00E-129$	0
Unigene47210_All	foxi1c	*Caenorhabditis brenneri*	$4.00E-16$	0
Unigene49637_All	OAR	*Crassostrea gigas*	$3E-96$	0
Unigene49559_All	notch-like protein 1	*Crassostrea gigas*	$3.00E-67$	0
CL26075.Contig2_All	Cbx1	*Mus musculus*	$5.00E-60$	0
CL4780.Contig2_All	XHOX-3	*Crassostrea gigas*	$1E-124$	0
Unigene23054_All	ZF E-box-binding		$4.00E-60$	0
Unigene30958_All	ZF protein 64		$6.00E-08$	0
Unigene18460_All	aristaless-like 4		$1E-120$	0
Unigene22708_All	ARX		$4.00E-87$	0
Unigene22940_All	IRX-6		$2E-101$	0
Unigene27129_All	engrailed		$2.00E-56$	0
Unigene27119_All	homeobox-like protein		$2.00E-45$	0
Unigene27517_All	SIX3		$4E-129$	0
Unigene31497_All	ceh-9		$4.00E-64$	0
Unigene31654_All	Slou		$1E-103$	0
Unigene40727_All	unc-4-like protein		$7E-139$	0
Unigene40815_All	homeobox 2		$7E-120$	0
Unigene49664_All	even-skipped-like protein 1		$2E-107$	0
Unigene50280_All	not2		$9E-88$	0
Unigene23318_All	engrailed-2-B		$8E-61$	3
Unigene51601_All	BarH-like 1		$6E-31$	3
Unigene31090_All	PKNOX2		$3E-156$	3
Unigene49991_All	HMX1		$2E-27$	3
Unigene41009_All	Pax-7		$4E-128$	3
Unigene17191_All	Pax-2-A		$5E-102$	3
CL11027.Contig2_All	Polycomb BMI-1-A	*Danio rerio*	$3.00E-78$	3
CL13897.Contig2_All	Polycomb suz12	*Xenopus tropicalis*	$2.00E-149$	3
CL20556.Contig2_All	pcgf3	*Xenopus tropicalis*	$1.00E-77$	3
CL26777.Contig2_All	EPC1	*Homo sapiens*	$4.00E-147$	3
CL15780.Contig2_All	Wnt-4	*Homo sapiens*	$3.00E-93$	3
Unigene44847_All	BMP 3/3b	*Branchiostoma japonicus*	$2.00E-59$	3
CL11806.Contig6_All	FOXP1	*Homo sapiens*	$5.00E-104$	3
Unigene32767_All	foxe1	*Crassostrea gigas*	$3.00E-81$	3
CL4477.Contig6_All	Msx2	*Crassostrea gigas*	$1.00E-22$	3
Unigene45481_All	ceh-37		$2E-59$	12
Unigene55326_All	Pax-8		$7E-59$	20
Unigene18153_All	corepressor 1-like		0	21
Unigene45344_All	SIX4		$7E-84$	21
Unigene45422_All	aristaless		$5E-72$	21
Unigene45788_All	HMX3-B		$1E-55$	21
Unigene40909_All	odd-skipped-related 1	*Crassostrea gigas*	$7.00E-74$	21
Unigene22645_All	Hox5	*Haliotis rufescens*	$3.00E-85$	24
Unigene35405_All	EGF-like 1	*Crassostrea gigas*	$6.00E-66$	24
Unigene17944_All	foxc2	*Crassostrea gigas*	$8.00E-174$	24

Table 5: Continued.

Unigene ID	Annotated gene	Reference species	*E* value	Trend
Unigene45321_All	Dorsal root ganglia		$2E-123$	26
CL664.Contig4_All	MAPK	*Homo sapiens*	$7.00E-127$	26
Unigene23113_All	LIM		$7E-122$	27
Unigene44988_All	DBX1-A		$3E-93$	27
CL19911.Contig2_All	foxg1	*Xenopus laevis*	$5.00E-25$	27
Unigene17797_All	Nkx-2.2a		$2E-137$	28
Unigene44641_All	Nkx-6.2		$1E-125$	28
Unigene22323_All	hhex		$3E-92$	28
Unigene22601_All	Xlox	*Euprymna scolopes*	$5.00E-55$	28
Unigene33207_All	GBX-1		$3.00E-07$	28
Unigene36590_All	HOX3		$8E-89$	28
Unigene45891_All	MOX-2		$9E-82$	28
CL20549.Contig2_All	bicaudal	*Xenopus laevis*	$1.00E-45$	28
Unigene23139_All	hedgehog	*Crassostrea gigas*	$5.00E-99$	28
CL13232.Contig2_All	EGF-like D1044.2	*Caenorhabditis elegans*	$9.00E-06$	28
CL178.Contig1_All	EGFR	*Apis mellifera*	0	28
CL5633.Contig3_All	MAPKAPK5	*Homo sapiens*	$5.00E-132$	28
Unigene22098_All	O-fut1	*Crassostrea gigas*	$1.00E-131$	28
Unigene31699_All	foxl2	*Crassostrea gigas*	$1.00E-119$	28
Unigene36180_All	foxl1	*Crassostrea gigas*	$1.00E-119$	28
Unigene40504_All	foxslp2	*Crassostrea gigas*	$8.00E-116$	28
Unigene31565_All	bmp 2b	*Crassostrea gigas*	$6.00E-96$	29
CL7953.Contig2_All	Notch	*Crassostrea gigas*	$1.00E-139$	29
Unigene36286_All	Notch-like protein 2	*Crassostrea gigas*	$5.00E-65$	29
Unigene27672_All	ZF C2H2	*Brugia malayi*	$4.00E-07$	29
Unigene27337_All	LOX2		$8.00E-98$	29
Unigene27686_All	BarH-like 1-like	*Oreochromis niloticus*	$3.00E-36$	29
CL12322.Contig2_All	ALDH16A1	*Bos taurus*	$1.00E-179$	29
CL3565.Contig3_All	ALDH2	*Crassostrea gigas*	0	29
CL1306.Contig2_All	Wnt 1	*Homo sapiens*	$7.00E-13$	29

related to immunity, indicating that innate protection is vital in the early developmental stages.

Differential gene expressions (DEGs) occurred mainly during early stage transitions (Figure 6). Most genes were up- or downregulated from trochophore to D-shaped and from D-shaped to umbonal stages, indicating that processes associated with these transitions are very complicated. Principal component analyses yielded consistent results, where we identified 10 unigenes attributed to the divergence among trochophore, D-shaped, and UES (umbonal, eyespots, and spats) stages in *P. fucata*, being highly expressed in the trochophore stage. However, no functional annotations match these functionally important sequences, indicating that further research would help to elucidate the molecular mechanism of metamorphosis in this species in the future.

The analysis of expression trends indicated that 12,009 of 13,277 unigenes are sorted into eight significant expression trend groups. Among the significant trends, there were 10,031 (trends 3, 0, and 2), 11,978 (trends 28, 29, 21, 24, 26, and 27), 9,046 (trend 28, 29, 26, and 27), 9,518 (trend 28, 29, 24, and 27), and 5,214 (trend 29, 24, and 26) unigenes displaying increased expression in trochophore, D-shaped, umbonal, eyespots, and spats stages, respectively. This conveys that more genes are expressed in the early stages, consistent with the DEG and PCA analyses in our study. Particularly, 6,653 unigenes (trend 3) were highly expressed only in the trochophore stage, 3,340 unigenes (trend 29) expressed in an increasing pattern over the course of development, and 2,631 unigenes (trend 0) expressed in a decreasing pattern. These genes are worth further investigation.

The KEGG pathway enrichment analysis indicated that most unigenes in trend 3 were involved in pathways of spliceosome or RNA transport, indicating that, in the early stage of *P. fucata*, RNA synthesis is more predominant. On the contrary, genes in trend 29 showed significant enrichment in translational elongation pathways, suggesting that protein synthesis is more and more prevalent during larval development. In trends 27 and 28, a large number of unigenes were involved in the calcium signaling pathway, synchronizing with the shell formation of prodissoconchs I and II in D-shaped and umbonal stages [24, 41]. In addition, immune pathways were also enriched, indicating that innate

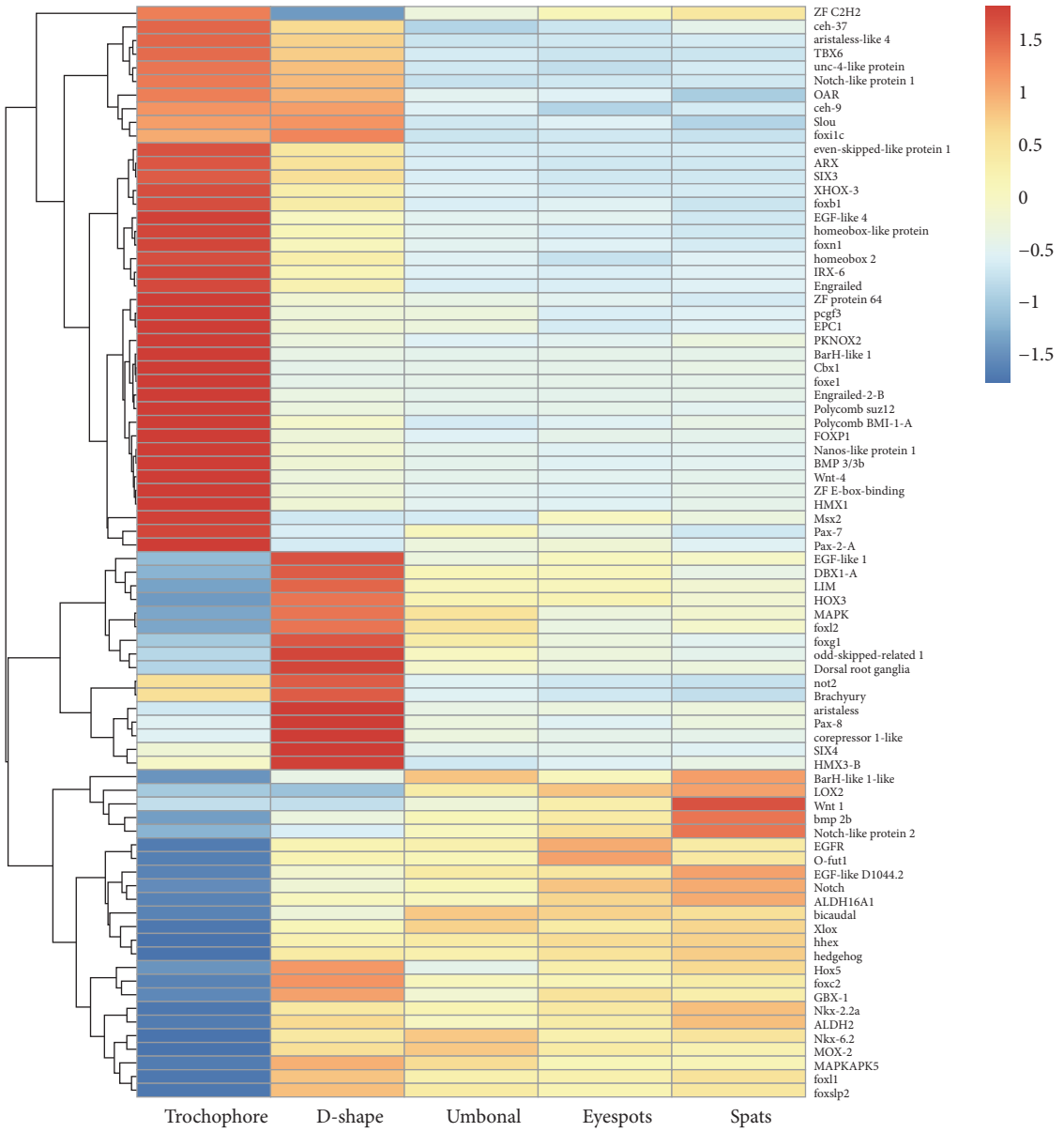


Figure 9: Clusters of expression patterns of development-related genes across five larval stages in *Pinctada fucata*.

protection is important during the entire course of larval development [3, 42, 43].

In our study, 80 known development-related, differentially expressed unigenes were identified throughout the five larval stages (Table 5). Half of them were homeobox-containing genes, including genes known to be involved in the development of body patterning (*engrailed*, *SIX3*, *Pax-7*, *LIM*, and *Hox* family members), suggesting that these genes play important roles in the metamorphic changes of *P. fucata* larvae. Nearly half of the homeobox-containing genes were upregulated in trochophore and D-shaped stages (Figure 9). We identified two *Hox* genes, *Hox5* and *Hox3* (Figure 9), which are highly expressed in D-shaped veliger, indicating that they are involved in the growth of D-shaped larvae. We also found early developmentally relevant signaling molecules such as Hedghog, TGFβ, and Wnt family, which are known to play important roles in axis formation, muscle differentiation, and nervous system development [26]. Recent evidence has suggested that classic morphogens, such as Wnts, TGFβ/BMP family members, and Hedgehogs, may all serve as axon guidance cues for a variety of axons in different organisms [44]. Several studies have provided increasing evidence that *Sonic hedgehog* (*Shh*) is an important axon guidance cue throughout vertebrate neural development [45, 46]. In our study, one hedgehog gene (Unigene23139_All) was identified and highly expressed in umbonal and eyespots stages (Figure 9), suggesting that increased neural development was likely taking place during those stages.

The Wnt signal pathway has been shown to play an important role in the segmentation of the marine polychaete

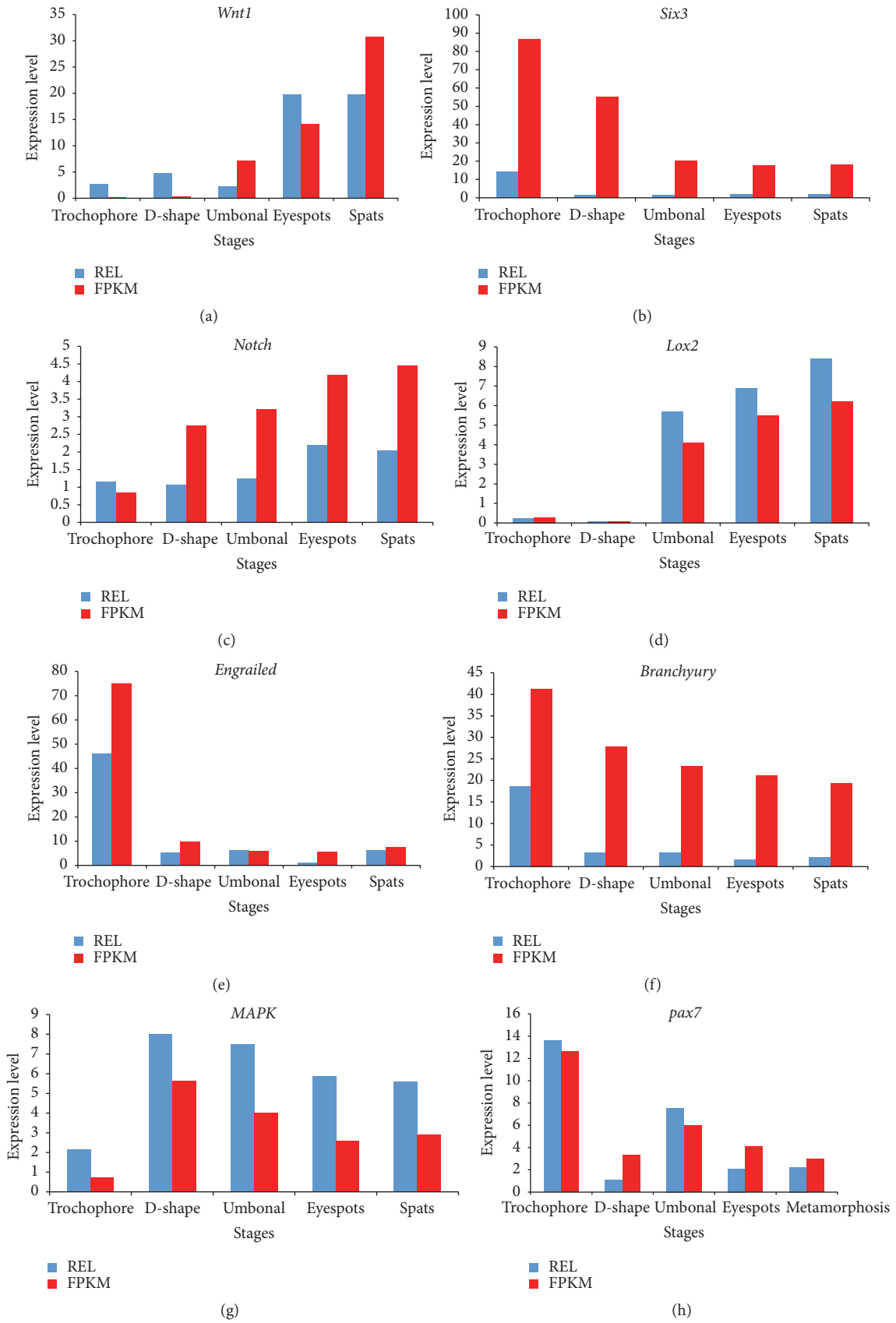


Figure 10: Expression changes in (a) *Wnt1*, (b) *Six3*, (c) *notch*, (d) *lox2*, (e) *engrailed*, (f) *branchyury*, (g) *MAPK*, and (h) *pax7* in trochophore, D-shaped, umbonal, eyespots, and spats larval stages by qPCR.

Capitella capitata [47, 48], while the maintenance of primitive hematopoiesis has been attributed to *Wnt4* in the vertebrate embryo [49]. Both *Wnt4* and *Wnt1* were observed in this study; *Wnt4* was highly expressed in the trochophore stage, while *Wnt1* was expressed in an increasing pattern over the course of larval development, suggesting possible involvement in blood formation from the beginning of development and in body transformation during all stages. Ten classes of fox genes were also found in our study, comprising the largest number of genes identified in the DEGs and displaying different trends in expression. *FoxL2*, XX-dominantly expressed in the differentiating ovaries of mammals [50], birds [51], and fish [52–54], was expressed highly in the D-shaped stage, suggesting the possible beginning of sexual development. Some important growth-related genes were also identified and differentially expressed among the five development stages, including *EGF-like*, *MAPK*, and *MAPKK* genes, which were actively expressed during the five developmental stages, and may contribute significantly to the transitions between developmental stages in *P. fucata* larvae.

Nonetheless, the body form transformations that take place during larval development involve a series of morphological and physiological changes and corresponding molecular changes, which have not been systematically studied and remain unclear. Therefore, a more broad understanding of the molecular underpinnings of important biological processes still merits further investigation.

5. Conclusions

In this study, a total of 174,126 unique transcripts were assembled and 60,999 were annotated. The number of unigenes varied between the five larval stages. The expression profiles of trochophore, D-shaped, and UES (umbonal, eyespots, and spats) larvae were distinctly different. Most unigenes were up- or downregulated in early stage transitions and 29 expression trends were sorted, eight of which were significant. In total, 80 development-related, differentially expressed unigenes were identified and eight were verified by qPCR. These observations should be helpful in understanding the molecular mechanisms of the larval metamorphic development of *P. fucata*.

Additional Points

Highlights.(i) A large number of assembled transcripts from *Pinctada fucata* are reported for the first time. (ii) Large variations in expression of DEGs related to development were observed in early larval stages. (iii) Twenty-nine expression trends were identified for the first time.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Haimei Li and Dahui Yu conceived and designed the project. Haimei Li, Bo Zhang, and Baosuo Liu cultured and collected the *P. fucata* larval samples and the adult samples. Haimei Li, Guiju Huang, and Sigang Fan carried out the transcriptome analysis and qPCR experiments. Dongling Zhang provided technical assistance. Haimei Li and Dahui Yu wrote the manuscript. All listed authors have read, edited, and approved the final manuscript.

Acknowledgments

This work was supported by The National Natural Science Foundation of China (NSFC) (31372525), the Earmarked Fund for China Agriculture Research System (Grant no. CARS-48), and Special Fund for Marine Fisheries Research and Extension of Guangdong Province (A201301A02, A201301A08, Z2014003 Z2014006, and Z2015009). The authors are also grateful to the Guangzhou Gene denovo Biotechnology Co., Ltd., for assisting in the data analysis.

References

[1] P.-Y. Qian, "Larval settlement of polychaetes," *Hydrobiologia*, vol. 402, pp. 239–253, 1999.

[2] T. Fujimura, K. Wada, and T. Iwaki, "Development and morphology of the pearl oyster larvae, *Pinctada fucata*," *Venus*, vol. 54, pp. 25–48, 1995.

[3] A. Heyland and L. L. Moroz, "Signaling mechanisms underlying metamorphic transitions in animals," *Integrative and Comparative Biology*, vol. 46, no. 6, pp. 743–759, 2006.

[4] C. Devine, V. F. Hinman, and B. M. Degnan, "Evolution and developmental expression of nuclear receptor genes in the ascidian *Herdmania*," *The International Journal of Developmental Biology*, vol. 46, no. 4, pp. 687–692, 2002.

[5] J. M. Arnold, R. Eri, B. M. Degnan, and M. F. Lavin, "Novel gene containing multiple epidermal growth factor-like motifs transiently expressed in the papillae of the ascidian tadpole larvae," *Developmental Dynamics*, vol. 210, no. 3, pp. 264–273, 1997.

[6] A. Nakayama, Y. Satou, and N. Satoh, "Isolation and characterization of genes that are expressed during *Ciona intestinalis* metamorphosis," *Development Genes and Evolution*, vol. 211, no. 4, pp. 184–189, 2001.

[7] R. Eri, J. M. Arnold, and V. F. Hinman, "Patterning of dopaminergic neurotransmitter identity among *Caenorhabditis elegans* ray sensory neurons by a TGF family signaling pathway and a Hox gene," *Development*, vol. 126, no. 24, pp. 5819–5831, 1999.

[8] B. M. Degnan and D. E. Morse, "Identification of eight homeobox-containing transcripts expressed during larval development and at metamorphosis in the gastropod mollusc *Haliotis rufescens*," *Molecular Marine Biology and Biotechnology*, vol. 2, no. 1, pp. 1–9, 1993.

[9] B. M. Degnan, S. M. Degnan, G. Fentenany, and D. E. Morse, "A Mox homeobox gene in the gastropod mollusc *Haliotis rufescens* is differentially expressed during larval morphogenesis and metamorphosis," *FEBS Letters*, vol. 411, no. 1, pp. 119–122, 1997.

[10] A. F. Giusti, V. F. Hinman, S. M. Degnan, B. M. Degnan, and D. E. Morse, "Expression of a *Scr*/*Hox5* gene in the larval central nervous system of the gastropod *Haliotis*, a non-segmented spiralian lophotrochozoan," *Evolution and Development*, vol. 2, no. 5, pp. 294–302, 2000.

[11] S. L. Coon, W. K. Fitt, and D. B. Bonar, "Competence and delay of metamorphosis in the Pacific oyster *Crassostrea gigas*," *Marine Biology*, vol. 106, no. 3, pp. 379–387, 1990.

[12] C. Lelong, M. Mathieu, and P. Favrel, "Structure and expression of mGDF, a new member of the transforming growth factor-β superfamily in the bivalve mollusc *Crassostrea gigas*," *European Journal of Biochemistry*, vol. 267, no. 13, pp. 3986–3993, 2000.

[13] T. J. Fiedler, A. Hudder, S. J. McKay et al., "The transcriptome of the early life history stages of the California sea hare *Aplysia californica*," *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, vol. 5, no. 2, pp. 165–170, 2010.

[14] J. D. Lambert, X. Y. Chan, B. Spiecker, and H. C. Sweet, "Characterizing the embryonic transcriptome of the snail *Ilyanassa*," *Integrative and Comparative Biology*, vol. 50, no. 5, pp. 768–777, 2010.

[15] P. Huan, H. Wang, and B. Liu, "Transcriptomic analysis of the clam *Meretrix meretrix* on different larval stages," *Marine Biotechnology*, vol. 14, no. 1, pp. 69–78, 2012.

[16] Z.-X. Huang, Z.-S. Chen, C.-H. Ke et al., "Pyrosequencing of *Haliotis diversicolor* transcriptomes: insights into early developmental molluscan gene expression," *PLoS ONE*, vol. 7, no. 12, Article ID e51279, 2012.

[17] J. Qin, Z. Huang, J. Chen, Q. Zou, W. You, and C. Ke, "Sequencing and de novo analysis of *Crassostrea angulata* (Fujian Oyster) from 8 different developing phases using 454 GSFLx," *PLoS ONE*, vol. 7, no. 8, Article ID e43653, 2012.

[18] S. Bassim, A. Tanguy, B. Genard, D. Moraga, and R. Tremblay, "Identification of *Mytilus edulis* genetic regulators during early development," *Gene*, vol. 551, no. 1, pp. 65–78, 2014.

[19] T. Takeuchi, T. Kawashima, R. Koyanagi et al., "Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology," *DNA Research*, vol. 19, no. 2, pp. 117–130, 2012.

[20] Y. Shi, C. Yu, Z. Gu, X. Zhan, Y. Wang, and A. Wang, "Characterization of the pearl oyster (*Pinctada martensii*) mantle transcriptome unravels biomineralization genes," *Marine Biotechnology*, vol. 15, no. 2, pp. 175–187, 2013.

[21] A. Wang, Y. Wang, Z. Gu, S. Li, Y. Shi, and X. Guo, "Development of expressed sequence tags from the pearl oyster, *Pinctada martensii* Dunker," *Marine Biotechnology*, vol. 13, no. 2, pp. 275–283, 2011.

[22] X. Zhao, Q. Wang, Y. Jiao et al., "Identification of genes potentially related to biomineralization and immunity by transcriptome analysis of pearl sac in pearl oyster *Pinctada martensii*," *Marine Biotechnology*, vol. 14, no. 6, pp. 730–739, 2012.

[23] S. Kinoshita, N. Wang, H. Inoue et al., "Deep sequencing of ESTs from nacreous and prismatic layer producing tissues and a screen for novel shell formation-related genes in the pearl oyster," *PLoS ONE*, vol. 6, no. 6, Article ID e21238, 2011.

[24] Y. Miyazaki, T. Nishida, H. Aoki, and T. Samata, "Expression of genes responsible for biomineralization of *Pinctada fucata* during development," *Comparative Biochemistry and Physiology—B: Biochemistry and Molecular Biology*, vol. 155, no. 3, pp. 241–248, 2010.

[25] J. Liu, D. Yang, S. Liu et al., "Microarray: a global analysis of biomineralization-related gene expression profiles during larval development in the pearl oyster, *Pinctada fucata*," *BMC Genomics*, vol. 16, no. 1, article 325, 2015.

[26] D. H. E. Setiamarga, K. Shimizu, J. Kuroda et al., "An *in-silico* genomic survey to annotate genes coding for early development-relevant signaling molecules in the pearl oyster, *Pinctada fucata*," *Zoological Science*, vol. 30, no. 10, pp. 877–888, 2013.

[27] H. Koga, N. Hashimoto, D. G. Suzuki et al., "A genome-wide survey of genes encoding transcription factors in Japanese pearl oyster *Pinctada fucata*: II. Tbx, Fox, Ets, HMG, NFκB, bZIP, and C2H2 zinc fingers," *Zoological Science*, vol. 30, no. 10, pp. 858–867, 2013.

[28] Y. Morino, K. Okada, M. Niikura, M. Honda, N. Satoh, and H. Wada, "A genome-wide survey of genes encoding transcription factors in the Japanese pearl oyster, *Pinctada fucata*: I. Homeobox genes," *Zoological Science*, vol. 30, no. 10, pp. 851–857, 2013.

[29] G. Pertea, X. Huang, F. Liang et al., "TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets," *Bioinformatics*, vol. 19, no. 5, pp. 651–652, 2003.

[30] C. Iseli, C. V. Jongeneel, and P. Bucher, "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences ," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, pp. 138–148, Heidelberg, Germany, 1999.

[31] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[32] J. Ye, L. Fang, H. Zheng et al., "WEGO: a web tool for plotting GO annotations," *Nucleic Acids Research*, vol. 34, pp. W293–W297, 2006.

[33] R. Li, C. Yu, Y. Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[34] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[35] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, article 191, 2006.

[36] M. Kanehisa, M. Araki, S. Goto et al., "KEGG for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, no. 1, pp. D480–D484, 2008.

[37] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[38] X.-D. Huang, M. Zhao, W.-G. Liu et al., "Gigabase-scale transcriptome analysis on four species of pearl oysters," *Marine Biotechnology*, vol. 15, no. 3, pp. 253–264, 2013.

[39] E. Meyer, G. V. Aglyamova, S. Wang et al., "Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx," *BMC Genomics*, vol. 10, article 219, pp. 1–8, 2009.

[40] R. Bettencourt, M. Pinheiro, C. Egas et al., "High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*," *BMC Genomics*, vol. 11, article 559, 2010.

[41] D. Fang, G. Xu, Y. Hu, C. Pan, L. Xie, and R. Zhang, "Identification of genes directly involved in shell formation and their functions in pearl oyster, *Pinctada fucata*," *PLoS ONE*, vol. 6, no. 7, Article ID e21860, 2011.

[42] E. A. Williams, B. M. Degnan, H. Gunter, D. J. Jackson, B. J. Woodcroft, and S. M. Degnan, "Widespread transcriptional changes pre-empt the critical pelagic-benthic transition in the vetigastropod *Haliotis asinina*," *Molecular Ecology*, vol. 18, no. 5, pp. 1006–1025, 2009.

[43] D. E. Clapham, "Calcium signaling," *Cell*, vol. 131, no. 6, pp. 1047–1058, 2007.

[44] F. Charron and M. Tessier-Lavigne, "Novel brain wiring functions for classical morphogens: a role as graded positional cues in axon guidance," *Development*, vol. 132, no. 10, pp. 2251–2262, 2005.

[45] F. Charron, E. Stein, J. Jeong, A. P. McMahon, and M. Tessier-Lavigne, "The morphogen sonic hedgehog is an axonal chemoattractant that collaborates with netrin-1 in midline axon guidance," *Cell*, vol. 113, no. 1, pp. 11–23, 2003.

[46] J. K. Chen, J. Taipale, M. K. Cooper, and P. A. Beachy, "Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened," *Genes & Development*, vol. 16, no. 21, pp. 2743–2748, 2002.

[47] E. C. Seaver and L. M. Kaneshige, "Expression of 'segmentation' genes during larval and juvenile development in the polychaetes *Capitella* sp. I and *H. elegans*," *Developmental Biology*, vol. 289, no. 1, pp. 179–194, 2006.

[48] J. L. Christian, B. J. Gavin, A. P. McMahon, and R. T. Moon, "Isolation of cDNAs partially encoding four *Xenopus* Wnt-1int-1-related proteins and characterization of their transient expression during embryonic development," *Developmental Biology*, vol. 143, no. 2, pp. 230–234, 1991.

[49] H. T. Tran, B. Sekkali, G. Van Imschoot, S. Janssens, and K. Vleminckx, "Wnt/β-catenin signaling is involved in the induction and maintenance of primitive hematopoiesis in the vertebrate embryo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 37, pp. 16160–16165, 2010.

[50] D. Baron, F. Batista, J. S. Chaffaux Cocquet et al., "*Foxl2* gene and the development of the ovary: a story about goat, mouse, fish and woman," *Reproduction Nutrition Development*, vol. 45, no. 6, pp. 729–729, 2005.

[51] M. S. Govoroun, M. Pannetier, E. Pailhoux et al., "Isolation of chicken homolog of the *FOXL2* gene and comparison of its expression patterns with those of aromatase during ovarian development," *Developmental Dynamics*, vol. 231, no. 4, pp. 859–870, 2004.

[52] D. Baron, J. Cocquet, X. Xia, M. Fellous, Y. Guiguen, and R. A. Veitia, "An evolutionary and functional analysis of *FoxL2* in rainbow trout gonad differentiation," *Journal of Molecular Endocrinology*, vol. 33, no. 3, pp. 705–715, 2004.

[53] M. Nakamoto, M. Matsuda, D.-S. Wang, Y. Nagahama, and N. Shibata, "Molecular cloning and analysis of gonadal expression of *Foxl2* in the medaka, *Oryzias latipes*," *Biochemical and Biophysical Research Communications*, vol. 344, no. 1, pp. 353–361, 2006.

[54] D.-S. Wang, T. Kobayashi, L.-Y. Zhou et al., "Foxl2 up-regulates aromatase gene transcription in a female-specific manner by binding to the promoter as well as interacting with Ad4 binding protein/steroidogenic factor 1," *Molecular Endocrinology*, vol. 21, no. 3, pp. 712–725, 2007.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2016, Article ID 9837310, 10 pages
http://dx.doi.org/10.1155/2016/9837310

Research Article

RNA Sequencing of Formalin-Fixed, Paraffin-Embedded Specimens for Gene Expression Quantification and Data Mining

Yan Guo,[1] Jie Wu,[2] Shilin Zhao,[1] Fei Ye,[3] Yinghao Su,[2] Travis Clark,[4] Quanhu Sheng,[1] Brian Lehmann,[5] Xiao-ou Shu,[2] and Qiuyin Cai[2]

[1]*Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA*
[2]*Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA*
[3]*Department of Biostatistics, Vanderbilt University, Nashville, TN, USA*
[4]*Genentech, Baltimore, MD, USA*
[5]*Department of Biochemistry, Vanderbilt University, Nashville, TN, USA*

Correspondence should be addressed to Qiuyin Cai; qiuyin.cai@vanderbilt.edu

Received 24 May 2016; Accepted 6 September 2016

Academic Editor: Brian Wigdahl

Background. Proper rRNA depletion is crucial for the successful utilization of FFPE specimens when studying gene expression. We performed a study to evaluate two major rRNA depletion methods: Ribo-Zero and RNase H. RNAs extracted from 4 samples were treated with the two rRNA depletion methods in duplicate and sequenced ($N = 16$). We evaluated their reducibility, ability to detect RNA, and ability to molecularly subtype these triple negative breast cancer specimens. *Results*. Both rRNA depletion methods produced consistent data between the technical replicates. We found that the RNase H method produced higher quality RNAseq data as compared to the Ribo-Zero method. In addition, we evaluated the RNAseq data generated from the FFPE tissue samples for noncoding RNA, including lncRNA, enhancer/super enhancer RNA, and single nucleotide variation (SNV). We found that the RNase H is more suitable for detecting high-quality, noncoding RNAs as compared to the Ribo-Zero and provided more consistent molecular subtype identification between replicates. Unfortunately, neither method produced reliable SNV data. *Conclusions*. In conclusion, for FFPE specimens, the RNase H rRNA depletion method performed better than the Ribo-Zero. Neither method generates data sufficient for SNV detection.

1. Background

Formalin-fixed paraffin-embedded (FFPE) tissue is the most common method of tissue preparation used in clinics. FFPE preservation was developed to maintain morphology without any special considerations of preserving nucleic acids. Therefore, the difficulty of evaluating gene expression levels in FFPE samples remains one of the biggest disadvantages of FFPE preservation because the process of fixing the tissue samples and embedding them in paraffin often leads to RNA degradation and chemical modification. Furthermore, a nucleic acid can be cross-linked with a protein during the formalin fixation process, and most of the RNA isolated from FFPE tissues is highly degraded and reduced to a much lower yield than that of RNA isolated from the same amount of fresh tissues. To that end, RNA isolated from recently embedded tissues will be of better quality than RNA isolated from older embedded tissues. As a result, when amplifying RNA with oligo-dT primers, there is an overrepresentation of 3′ data due to the fragmented nature of RNA isolated from FFPE tissues.

Given the aforementioned reasons, gene expression analysis based on FFPE samples has been historically challenging. The most critical step in a FFPE sample based study is tissue preparation, as it ensures the integrity of the yield and data quality. It has been greatly emphasized that improper FFPE tissue preparation can diminish the quality of the nucleic acids from the tissue, limiting their use for gene expression profiling [1]. Yet, FFPE samples are often sought after due to their in-depth retrospective records. The success

of a FFPE sample based study often depends on several steps: RNA isolation, reverse transcription, qPCR primer design, and preamplification. With carefully designed preparation protocols, FFPE samples have been proven to be an invaluable source for gene expression studies. The potential applications of FFPE samples in biomedical research are substantial.

The vast majority of cellular RNA (>80%) is composed of noninformative ribosomal RNAs (rRNAs, 28 S, 5.8 S, and 18 S rRNAs) that require removal prior to cDNA synthesis for a RNA-seq library. For high-quality RNA samples, polyadenylated RNA is enriched from intact RNA using oligo-dT primers. Since the rRNA does not have a poly-A tail, it is removed prior to cDNA synthesis along with other informative, non-polyA RNA species. RNA samples isolated from FFPE tissues have two features that are not compatible with oligo-dT primer selection: fragmented RNA that produces $3'$ bias from oligo-dT selection, and, the degradation of the poly-A tail, thereby impacting the yield of recovered mRNA. Currently, there are two major rRNA depletion methods used for RNA isolated from FFPE samples: the Ribo-Zero rRNA removal kit (Epicentre/Illumina) and the RNase H method (also known as SDRNA) [2–4]. The Ribo-Zero kit uses a biotinylated antisense set of DNA capture probes that preferentially bind to rRNA. Magnetic beads are then used to capture the rRNA:DNA capture probe duplex. The resulting non-rRNA is left for cDNA synthesis. The RNase H method uses a similar initial depletion strategy by annealing 50–80 bp antisense DNA probes to the rRNA forming RNA:DNA hybrids. The RNA:DNA hybrids are treated with endoribonuclease RNase H that specifically degrades the phosphodiester bonds of RNA hybridized to DNA. This step is followed by a DNase I treatment to degrade the excess DNA probes. The resulting RNA is then ready for cDNA synthesis.

In the 2000s, microarray technology dominated high-throughput gene expression profiling but has since been replaced by RNAseq technology [5–9]. Successful gene expression studies based on FFPE samples by microarray technology [10–12] are much more abundant than studies using the relatively newer RNAseq technology. Here, we apply both RNA depletion methods, Ribo-Zero and RNase H, to isolated RNA from FFPE specimens to compare the overall qualities of data.

Furthermore, based on the premise that sequencing data offers exciting opportunities for additional data mining [13, 14, 16], we examined the data mining practicability of three types of supplementary information: SNVs, lncRNAs, and enhancer RNAs. SNVs are traditionally identified through DNA samples. SNV detection through RNAseq data has been historically challenging, although, with careful quality control, SNVs are detectable in RNAseq data [17–20]. Long noncoding RNAs (lncRNAs) are arbitrarily defined as longer than 200 nucleotides in length and do not encode proteins. Recent findings have suggested that lncRNAs play important roles in various diseases [21–28], and lncRNAs are detectable through the total RNAseq preparation method by the Ribo-Zero RNA rRNA removal kit [29]. Enhancer RNAs are a type of RNA that regulate spatiotemporal gene expression and impart cell-specific transcriptional outputs [30]. Recent advancements in RNAseq technology have enabled the ready detection of enhancer RNA [15, 30]. Super enhancer RNAs are a subset of enhancer RNA that are associated with cell identity and genetic risk of various diseases [31–33]. Our unique set of FFPE RNAseq data allows us to answer the question of whether a FFPE sample based RNAseq can be used for these types of data mining and determine which RNA isolation kit produces data most ideal for data mining.

2. Methods

2.1. Sample Description. To evaluate the practicability and effectiveness of gene expression profiling using FFPE samples, we designed a study using four triple negative breast cancer (TNBC) FFPE tumor tissue samples. The H&E slides were reviewed by a study pathologist and tumor tissues were dissected from an unstained FFPE tissue section for total RNA extraction. The tumor tissue sections were stored in a vacuum chamber at 4°C for eight to nine years before RNA isolation was performed. Total RNA was extracted and purified using a Qiagen's miRNeasy FFPE Kit, a kit specifically designed for purifying the total RNA and microRNA from FFPE tissue sections. The input RNA amount for both Ribo-Zero and RNase H rRNA depletion methods was 200 ng each. The quantity and quality of the RNA samples extracted from tumor tissue FFPE sections were checked by Nanodrop (E260, E260/E280 ratio, spectrum 220–320 nm) and by separation on an Agilent BioAnalyzer. Total RNA extracted from each of the four tumors was split into two samples (for a total of eight samples). Two rRNA depletion methods were used: Ribo-Zero and RNase H. Each of the eight samples was treated with the two rRNA depletion methods, prepared for library using TruSeq RNA sample Prep Kit v2 (Illumina), and sequenced by BGI Americas. In total, 16 RNAseq libraries were generated following manufacture protocols and sequenced on two lanes (for a total of eight samples per lane). The qualified libraries were amplified on cBots to generate the cluster on the flow cell. The amplified flow cell was sequenced paired-end on the HiSeq 2000 at read length of the 90 base pairs.

2.2. Data Processing. RNAseq data was thoroughly quality-controlled at multiple stages (raw, alignment, and expression) following the recommendation by Guo et al. [34]. Raw data and alignment were quality-controlled using QC3 [35], while expression data was quality-controlled using MultiRankSeq [36]. Alignments were performed using Tophat 2 [37] against the HG19 human reference genome. Read counts for protein coding RNAs, lncRNAs enhancer RNAs, and super enchanter RNAs for each sample were obtained using HTSeq [38] against the collective General Transfer Format (GTF) file build from Ensembl Human GTF v74, Gencode lncRNA v 1.9, and enhancer RNA coordinates provided in [15]. Read count data for each type of the RNA was normalized to the total read counts of each sample. Cluster analysis was performed using Heatmap3 to identify similarities among samples [39]. Spearman's correlation coefficients were used to denote the distance between any two samples.

Table 1: Sample description and alignment statistics.

ID	Library	Raw data			Alignment			
		Total reads	BQ	GC	CR	Non-CR	CR MQ	Non-CR MQ
1	Ribo-Zero	17.8 M	31	71.4%	27.7%	72.3%	32	47
2	Ribo-Zero	16.1 M	30	76.8%	31.4%	68.6%	23	47
3	Ribo-Zero	16.8 M	31	70.7%	31.6%	68.4%	28	46
4	Ribo-Zero	14.0 M	31	72.3%	46.0%	54.0%	34	47
1	Ribo-Zero	16.1 M	31	70.4%	27.4%	72.6%	31	47
2	Ribo-Zero	14.9 M	31	75.4%	37.9%	62.1%	21	47
3	Ribo-Zero	17.6 M	31	71.1%	29.8%	70.2%	29	47
4	Ribo-Zero	15.6 M	31	70.0%	46.1%	53.9%	36	47
1	RNase H	20.2 M	35	51.6%	79.9%	20.1%	46	33
2	RNase H	20.5 M	36	39.8%	42.6%	57.4%	45	41
3	RNase H	20.4 M	35	51.6%	78.9%	21.1%	45	33
4	RNase H	21.4 M	35	48.6%	58.0%	42.0%	45	37
1	RNase H	22.1 M	35	52.5%	80.3%	19.7%	46	35
2	RNase H	22.4 M	34	55.1%	74.7%	25.3%	44	33
3	RNase H	20.6 M	35	52.0%	78.5%	21.5%	45	31
4	RNase H	24.0 M	34	53.6%	80.1%	19.9%	45	30

CR: coding region; BQ: base quality; MQ: mapping quality; GC: GC content.

2.3. TNBC Subtype. Triple negative breast cancer (TNBC) is known to be molecularly and transcriptionally heterogeneous and can be classified into one of six subtypes (basal-like 1, BL1; basal-like 2, BL2; immunomodulatory, IM; mesenchymal, M; mesenchymal-stem like, MSL; and luminal AR, LAR) based on centroid correlations using gene expression [40]. In order to determine if RNAseq data originated from FFPE specimens can be used for clinical subtyping, we performed TNBC subtyping on each of the samples using TNBCtype [41] and compared the repeatability of TNBC subtyping consistency between the Ribo-Zero and RNase H methods.

2.4. NanoString. NanoString nCounter data was obtained on 302 genes using the same samples. The detailed processing and normalization method is described in [42]. We computed Spearman's correlation coefficients to evaluate the concordance between RNAseq and NanoString technology.

2.5. SNV Detection. We conducted advanced data mining on our FFPE RNAseq data to extract SNV. We inferred SNVs using Varscan 2 [43]. SNV quality was assessed by the transition/transversion (Ti/Tv) ratio and the pairwise heterozygous genotype consistency rate between any two samples. The Ti/Tv ratio is commonly used as a quality control measurement [44–46]. The Ti/Tv ratio of SNVs residing in coding regions should be between two and three and slightly lower for SNVs residing outside of the coding regions [47]. Higher Ti/Tv ratios, without exceeding the upper bound, usually indicate better overall quality. SNVs were annotated with ANNOVAR [48]. The heterozygous consistency rate of a pair of samples A-B is defined as the number of consistent genotypes between samples A and B, divided by the number of total heterozygous genotypes within B. A SNV is qualified as part of a consistency rate computation if it is detected by both samples and if the read depth for that SNV is at least 10 on both samples.

3. Results

3.1. Raw Data Quality Assessment. On average, the Ribo-Zero rRNA removal method produced 16.1 (range: 14.0–17.8) million reads per sample, and the RNase H produced 21.4 (range: 20.2–24.0) million reads per sample. The RNase H method consistently produced more reads than Ribo-Zero. Given that the same amount of RNA was used and the same number of samples was pooled per lane, a higher RNA capture efficiency is probable for RNase H than that of Ribo-Zero. On average, the guanine-cytosine (GC) content of Ribo-Zero was 72.3% (range: 70.0–76.8%), which was above the expected value (50%), whereas the GC content of the RNase H method was 50.6% (range: 39.8–55.1%). The GC content of the reference genome is roughly the expected GC content for the sequenced data. The GC content is 39.3% for the entire human genome, 48.9% for protein coding RNA, 39.7% for lncRNA, and 50.2% for rRNA. The sequenced reads of total RNAseq data are a mixture of protein coding RNA, lncRNA, and other species of RNA. With the expected GC content around 50%, RNase H produced data with GC content closest to the expected value. The raw data quality control only provided partial quality assessment of the samples.

3.2. Alignment Quality Assessment. Next, we examined the percentage of the reads that aligned to the coding region (Table 1). For the Ribo-Zero, on average, 34.7% (range: 27.4–46.1%) of the sequenced reads aligned to coding regions, and for the RNase H, on average, 71.6% (range: 42.6–80.3%) of the sequenced reads aligned to coding regions. An interesting

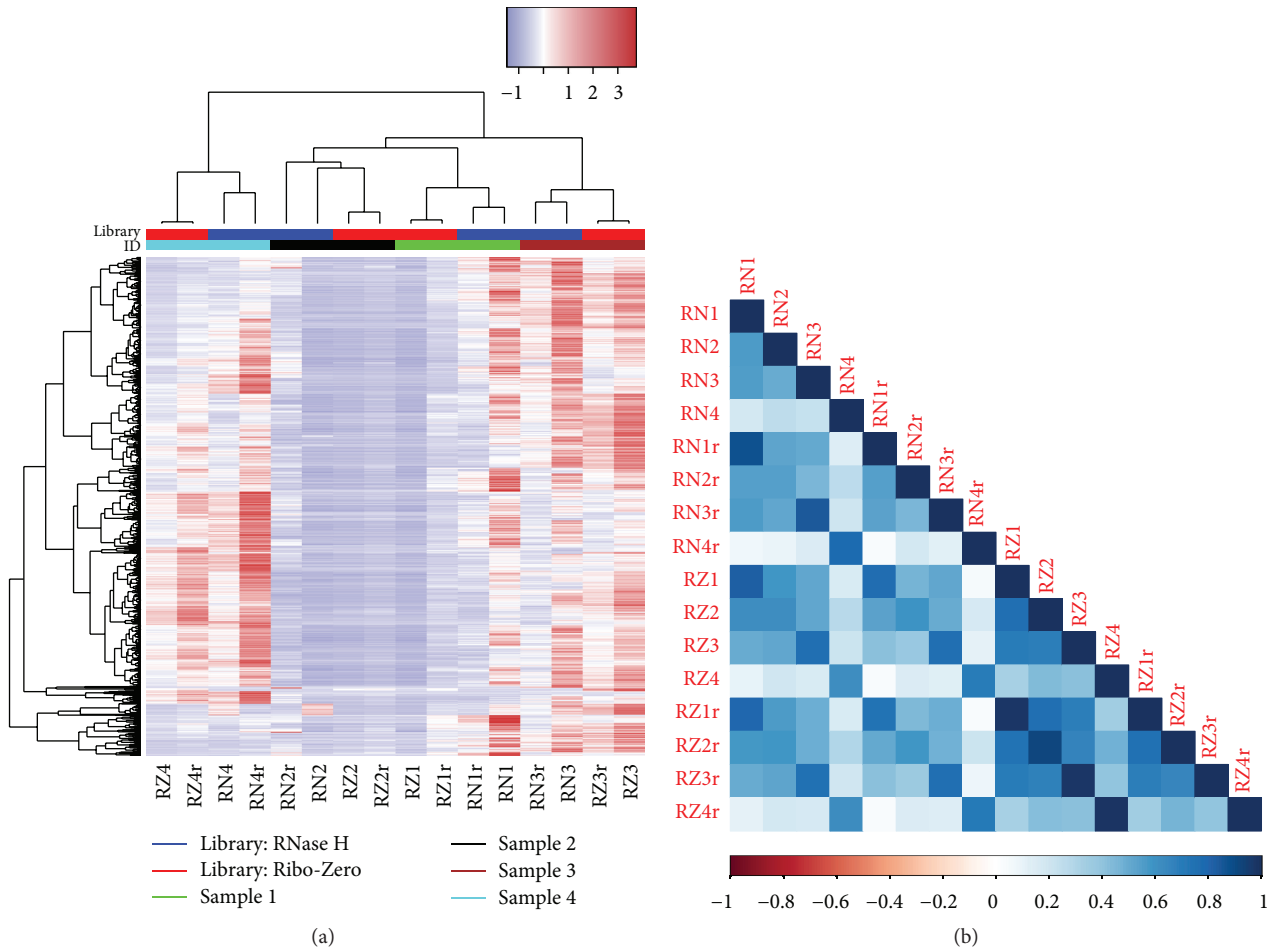


Figure 1: (a) Unsupervised cluster using all detected RNAs. Samples were clustered first by replicates then by rRNA depletion method. (b) Pairwise Spearman correlation heatmap between all samples. Ribo-Zero produced higher correlation between repeats than RNase H. The samples RN4 and RN4r produce low correlations with other samples compared to other random pairs. This could be the result of variation in the sample or variation introduced by the RNase H kit.

observation was made in regard to the mapping quality (MQ). Ribo-Zero produced higher mapping quality data in the noncoding region, whereas the RNase H method produced higher mapping quality in the coding region. For the Ribo-Zero, the average MQ for the coding region was 29 (range: 23–36) and 47 (range: 46-47) for the noncoding region. For the RNase H method, the average MQ was 45 (range: 44–46) for the coding region and 34 (range: 33–41) for the noncoding region. One of the repeats of sample two, which used the RNase H, is a potential outlier because it had the lowest GC content (39.8%) and the lowest coding region alignment rate (42.6%) of all the RNase H based samples. RNase H also produced less percentage of rRNA reads compared to Ribo-Zero (paired t-test $p = 0.03$).

3.3. Cluster Analysis. Cluster analysis showed that, regardless of which RNA isolation kit was used, the repeated sample clustered together based on gene expression. Within repeated samples, the rRNA depletion kits were clustered separately. The cluster analysis results provided additional evidence of quality concern for the RNase H sample two repeat one, as it was the only sample that did not perfectly cluster with its pair within the same RNA isolation kit (Figure 1(a)). The correlation heatmap (Figure 1(b)) showed similar results as presented in Figure 1(a). Essentially, we observed a higher pairwise correlation between repeated samples than between random samples.

3.4. TNBC Subtype Comparison. Overall, correlations to the TNBC subtypes were similar in replicates (Figure 2). RNase H samples had more consistent TNBC subtype calls between replicates (3/4 matching) than the Ribo-Zero samples (2/4 matching). The nonmatching replicate in the RNase H samples is sample 2 where we have previously noted its quality issue. This result suggests that RNase H produces RNAseq data with more consistent TNBC subtyping.

3.5. NanoString Comparison. We computed Spearman's correlation coefficients using the gene expression levels between RNAseq. The correlation dot plot (Figure 3) shows that the average correlation between Ribo-Zero and NanoString is 0.59 (range: 0.53–0.67), and the average correlation between

Sample	BL1	BL2	IM	M	MSL	LAR	Subtype
RN1	0.05	−0.28	0.34	−0.18	0.10	−0.04	MSL
RN1r	−0.03	−0.24	0.22	−0.13	0.18	0.01	MSL
RN2	−0.30	0.00	0.00	−0.08	0.24	0.25	LAR
RN2r	−0.12	0.12	0.09	−0.09	0.06	0.05	IM
RN3	0.34	−0.19	−0.18	0.23	−0.10	−0.11	LAR
RN3r	0.17	−0.17	−0.17	0.17	−0.01	−0.08	BL2
RN4	−0.18	0.33	−0.15	0.02	−0.01	0.09	LAR
RN4r	0.00	0.20	−0.13	0.08	−0.15	−0.08	IM
RZ1	−0.12	−0.17	0.08	−0.12	0.16	0.20	BL1
RZ1r	−0.18	−0.32	0.38	−0.32	0.25	0.16	BL1
RZ2	−0.15	0.10	0.14	−0.20	0.09	0.17	BL1
RZ2r	−0.21	0.22	0.26	−0.29	0.06	0.12	BL1
RZ3	0.25	−0.09	−0.16	0.14	−0.13	−0.14	BL2
RZ3r	0.24	−0.04	−0.12	0.14	−0.10	−0.12	BL2
RZ4	−0.12	0.29	−0.12	0.02	−0.15	0.01	BL2
RZ4r	−0.08	0.32	−0.08	0.02	−0.15	−0.05	BL2

BL1: basal-like 1
BL2: basal-like 2
IM: immunomodulatory
M: mesenchymal
MSL: mesenchymal-stem like
LAR: luminal AR

Figure 2: TNBC subtype results from TNBC type. The results show that RNase H samples produced better TNBC subtype consistency than Ribo-Zero samples.

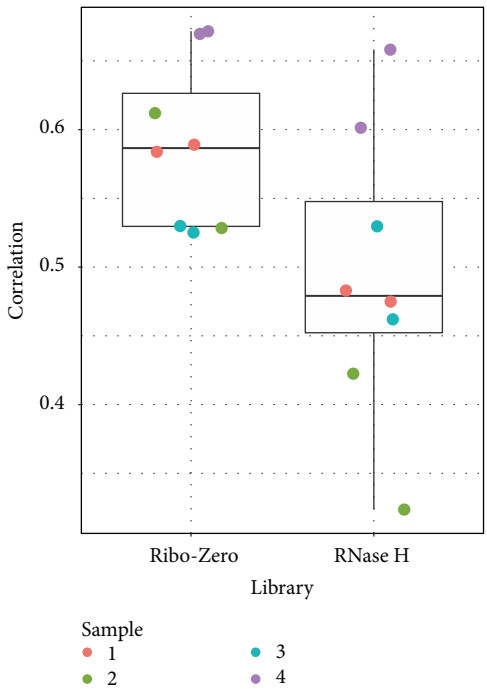


Figure 3: Spearman's correlation coefficients between RNAseq data and NanoString data. The Ribo-Zero samples produced slightly higher correlation with NanoString data then RNase H samples.

RNase H and NanoString is 0.49 (range: 0.32–0.66). The lowest correlation was produced by RNase H sample two repeat one which is likely to be a sample with a sequencing quality issue.

3.6. RNA Detection. We examined four kinds of RNAs: mRNA (Figure 4(a)), lncRNA (Figure 4(b)), enhancer RNA (Figure 4(c)), and super enhancer RNA (Figure 4(d)). After normalization by total read count, we used four detection thresholds (>0, >2, >5, and >10) to compare the RNA detection rates between the two RNA isolation kits. For all four types of RNAs, the Ribo-Zero rRNA depletion method detects more RNA at detection thresholds >0 and >2. When higher detection thresholds were used, the RNase H managed to detect more RNAs. RNA detected with low expression values could be the result of noises and is therefore less trustworthy than RNA detected with higher levels of expression. Based on these results, the RNase H rRNA depletion method detected more potentially reliable RNA as compared to the Ribo-Zero.

3.7. SNV Detection. We inferred SNVs from the FFPE RNA data using VarScan 2. After filtering for high quality SNVs (depth > 20), on average, the Ribo-Zero samples identified 525 SNVs per sample (range: 73–1862), and the RNase H samples identified 57747 SNVs per sample (range: 21932–87146). The RNase H samples clearly identified more SNVs than the Ribo-Zero prepared samples. This is caused by the difference of number of callable sites between the two kits. We defined a callable site to be a genomic position with coverage depth ≥ 20. RNase H produced substantially more

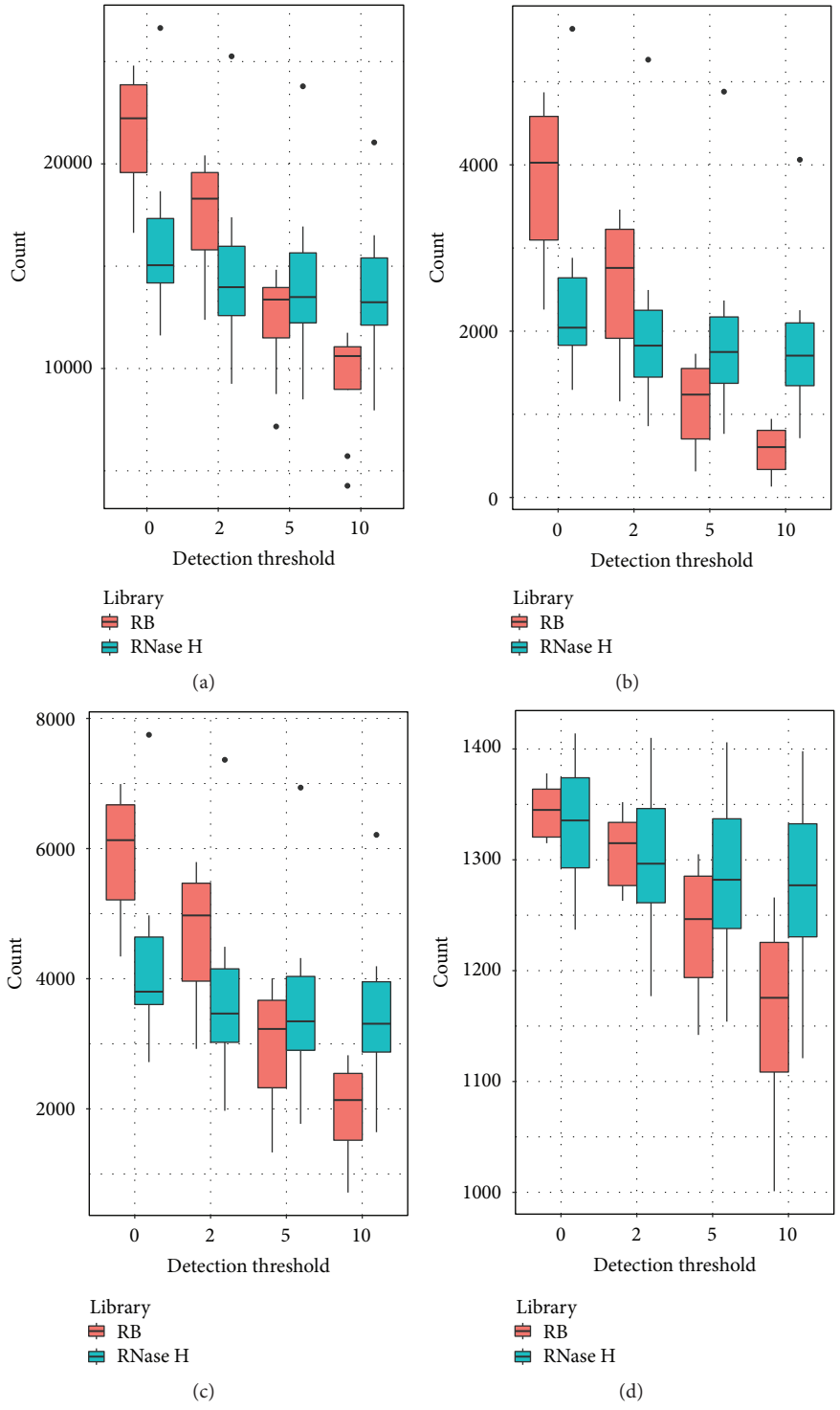


Figure 4: Detected RNA using thresholds: normalized reads count > 0, 2, 5, and 10. (a) Protein coding RNA. (b) lncRNA. (c) Enhancer RNA. (d) Super enhancer RNA. At lower thresholds (more noise), Ribo-Zero samples detected more RNAs. At higher thresholds (more reliability), RNase H method detected more RNAs.

callable sites than Ribo-Zero (Figure 5). The callable site analysis result shows that the coverage of Ribo-Zero is more spread out than RNase H. High variations in the number of SNVs were observed for both RNA isolation kits. For SNVs identified in coding regions, on average, the Ti/Tv ratio for Ribo-Zero was 3.51 (range: 2.42–8.00) and 2.08 (range: 1.34–2.47) for RNase H. For SNVs identified in noncoding regions, on average, the Ti/Tv ratio for Ribo-Zero was 2.84

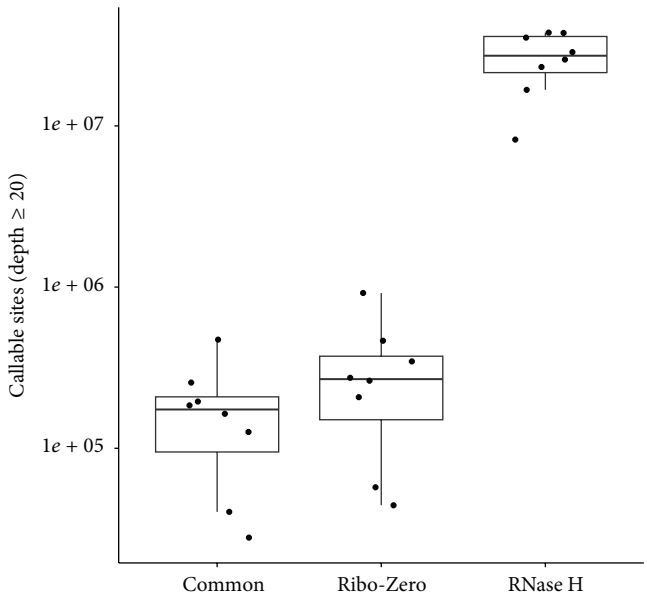


Figure 5: Callable site is defined as a genomic position with depth coverage ≥ 20. The number of callable sites indicates the number of genomic positions that are suitable for SNV inference. RNase H had substantially more callable sites than Ribo-Zero. The percentage of difference in callable site is significantly more than the percentage of difference in number of total reads sequenced by the two kits. *Y*-axis is plotted in $\log_{10}$ scale.

(range: 2.37–3.84) and 3.74 (range: 1.04–5.27) for RNase H. The variation for the Ti/Tv ratio is large, indicating potential problems with the SNVs identified.

Additional evidence for problematic SNV inferences was observed in the results of the pairwise heterozygous genotype consistency between samples. In DNA sequencing, we expect the heterozygous genotype consistency rate for technical replicates to be above 0.99. For RNAseq, the consistency rate is expected to be lower but still yield above 80%. However, on average, the consistency rates for both kits were less than 40% which were substantially below expectation. Restricting SNV pairwise heterozygous consistency computation to SNVs with depth greater than 50x for both samples in the pair increased the consistency slightly but still remained <50%. The low heterozygous consistency rates indicate that SNVs inferred from FFPE RNAseq samples contain high false positive rates and are therefore not ideal sources for detecting SNV.

4. Discussion

Utilization of FFPE specimens for gene expression studies could open a new avenue for molecular epidemiological and clinical research. Yet to date, the low quality of RNA from FFPE specimens for gene expression analysis has been a challenge. Several technologies have been developed for quantifying gene expression from FFPE specimens, such as NanoString [49] and quantitative Nuclease Protection Assay [50].

Since gene expression data can yield both molecular subtype classification and predicative markers of risk, efforts have been made to use RNA extracted from FFPE tissue on NanoString and microarray platforms [51, 52]. Triple negative breast cancer has been shown to be transcriptionally heterogeneous, with several molecular subtypes with differing biology [40, 53, 54]. The ability to identify TNBC subtypes from RNA isolated from FFPE tissues will provide opportunities for future clinical trial designs and retrospective evaluations of previously failed clinical trials by individual subtypes. To determine if RNA extracted from FFPE tissue that has been stored for eight to nine years could yield gene expression profiles by RNAseq sufficient enough to subtype TNBC, we compared the efficiencies of both the Ribo-Zero and the RNase H methods for rRNA depletion.

Through thorough quality control and analyses, we found that expression profiling of coding and noncoding RNA is possible for aged FFPE samples with RNAseq technology. The Ribo-Zero and RNase H method each had strengths and weaknesses in different areas. Our analyses suggested that RNase H is more suitable for studies that target protein coding RNA. On the other hand, Ribo-Zero offered more consistency between repeated samples, which is of pivotal importance, especially for low quality RNA extracted from FFPE tissues. Under the same amount of library input and same multiplexing scenario, RNase H consistently produced more reads than Ribo-Zero. Many reasons could have caused this read counts difference, including batch effect of the cluster on the flow cell, and library efficiencies. The evidences of more total reads sequenced under the same input amount and better rRNA depletion efficiency for RNase H support that RNase H has better library efficiency than Ribo-Zero. RNase H hybridizes directly to the sequences of rRNAs without the requirement of perfect match. The Ribo-Zero uses bait strategy which is similar to enrichment like exome capture with baits and beads. Thus it does not remove degraded, fragmented rRNAs as efficient as RNase H. Our study confirms previous finding that RNase H performed better than Ribo-Zero for low quality RNAs [55].

Furthermore, genes quantified from Ribo-Zero processed RNAseq data also had a slightly higher correlation with genes quantified by NanoString technology. This suggests that Ribo-Zero might offer better repeatability, although the correlation (50–60%, FFPE) with NanoString data (FFPE) did not reach the high correlation (80–90%, fresh frozen) between microarray and RNAseq [56]. We suspect this is primarily due to the variation introduced by the degraded quality of the RNA extracted from FFPE samples.

The subtyping of gene expression profiles obtained by both methods demonstrated that RNA isolated from stored FFPE samples can be used to determine distinct TNBC subtypes. While TNBC subtypes were similar among replicates, RNase H samples had more consistent TNBC subtype calls between replicates than that of the Ribo-Zero samples, which is potentially due to the more efficient capture of protein coding RNA.

By performing SNV detection analysis, we found that SNV detected by FFPE RNAseq data is subjected to quality concerns. It has been suggested that the SNV data inferred

from RNAseq data has a high false positive rate [57]. Several factors can contribute to the high false positive rate of SNV. First, alignment on RNAseq data can be more complicated than DNA sequencing data [19]. Processes such as RNA editing, alternative splicing, gene fusion, and polyadenylation introduce additional complications in RNAseq alignment. The step that reverse-transcribes RNA to cDNA can also introduce random errors. We have found that the number of SNVs inferred from RNAseq data can be several folds higher than that from the exome sequencing data on the sample. In our study, the lower quality of RNA isolated from FFPE tissue will result in an even higher number of false positive SNVs. The low consistency rate of SNVs identified between paired samples suggests that RNAseq data from FFPE tissues are not suitable for SNV inference.

5. Conclusion

Recent studies have shown remarkably high consistency between RNAseq data generated from paired freshly frozen and FFPE tissue samples [58–60]. Our study provides additional evidence for the practicability of conducting gene expression RNAseq with FFPE tissues. There is no denying that there are technical and quality limitations for FFPE RNAseq data. However, the majority of the issues can be overcome through thorough quality control and careful bioinformatics analyses. Our study supports the notions that RNAseq on FFPE samples can be used as an unbiased and comprehensive assessment of gene expression in biomedical studies, and RNase H method provides more efficient rRNA depletion than Ribo-Zero method for low quality fragmented RNAs.

Abbreviations

FFPE: Formalin-fixed paraffin-embedded
RNAseq: RNA sequencing
SNV: Single nucleotide variant
TNBC: Triple negative breast cancer.

Additional Points

Availability of Data and Materials. The sequencing data used in this study have been deposited into Gene Expression Omnibus (GEO) under the accession number GSE74270.

Ethical Approval

The study was approved by the institutional review board of Vanderbilt University.

Consent

All participants provided written informed consent during in-person interviews.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Yan Guo performed the sequencing analysis and wrote the manuscript. Shilin Zhao assisted with sequencing data analysis. Fei Ye performed the Nanostring data analysis. Jie Wu prepared the samples. Yinghao Su prepared the samples. Qiuyin Cai designed the study and contributed to the writing of the manuscript. Brian Lehmann performed the TNBC subtype analysis. Quanhu Sheng performed the TNBC subtype analysis and assisted with sequencing analysis. Travis Clark contributed to writing of the manuscript. Xiao-ou Shu designed the overall study.

Acknowledgments

Yan Guo was supported by P30 CA068485. We would also like to thank Stephanie Page Hoskins for editorial support. RNAseq and sample collection were supported by R01CA064277, R01CA118229, U01CA161045, and P50CA098131. RNA sample preparation was conducted at the Survey and Biospecimen Shared Resources, which is supported in part by the Vanderbilt-Ingram Cancer Center (P30CA068485).

References

[1] S. M. Hewitt, F. A. Lewis, Y. Cao et al., "Tissue handling and specimen preparation in surgical pathology: issues concerning the recovery of nucleic acids from formalin-fixed, paraffin-embedded tissue," *Archives of Pathology and Laboratory Medicine*, vol. 132, no. 12, pp. 1929–1935, 2008.

[2] X. Adiconis, D. Borges-Rivera, R. Satija et al., "Comparative analysis of RNA sequencing methods for degraded or low-input samples," *Nature Methods*, vol. 10, no. 7, pp. 623–629, 2014, Nature Methods, vol. 11, pp. 210, 2013.

[3] R. Huang, M. Jaritz, P. Guenzl et al., "An RNA-seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncrnas," *PLoS ONE*, vol. 6, no. 11, Article ID e27288, 2011.

[4] J. D. Morlan, K. Qu, and D. V. Sinicropi, "Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue," *PLoS ONE*, vol. 7, no. 8, Article ID e42882, 2012.

[5] Y. W. Asmann, E. W. Klee, E. A. Thompson et al., "3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer," *BMC Genomics*, vol. 10, article 531, 2009.

[6] N. Cloonan, A. R. R. Forrest, G. Kolle et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[7] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.

[8] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[9] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[10] X.-J. Ma, Z. Wang, P. D. Ryan et al., "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen," *Cancer Cell*, vol. 5, no. 6, pp. 607–616, 2004.

[11] L. Mittempergher, J. J. de Ronde, M. Nieuwland et al., "Gene expression profiles from formalin fixed paraffin embedded breast cancer tissue are largely comparable to fresh frozen matched tissue," *PLoS ONE*, vol. 6, no. 2, article e17163, 2011.

[12] P. T. Nelson, D. A. Baldwin, L. M. Scearce, J. C. Oberholtzer, J. W. Tobias, and Z. Mourelatos, "Microarray-based, high-throughput gene expression profiling of microRNAs," *Nature Methods*, vol. 1, no. 2, pp. 155–161, 2004.

[13] L. Han, K. C. Vickers, D. C. Samuels, and Y. Guo, "Alternative applications for distinct RNA sequencing strategies," *Briefings in Bioinformatics*, vol. 16, no. 4, pp. 629–639, 2014.

[14] D. C. Samuels, L. Han, J. Li et al., "Finding the lost treasures in exome sequencing data," *Trends in Genetics*, vol. 29, no. 10, pp. 593–599, 2013.

[15] G. Vahedi, Y. Kanno, Y. Furumoto et al., "Super-enhancers delineate disease-associated regulatory nodes in T cells," *Nature*, vol. 520, no. 7548, pp. 558–562, 2015.

[16] K. C. Vickers, L. A. Roteta, H. Hucheson-Dilks, L. Han, and Y. Guo, "Mining diverse small RNA species in the deep transcriptome," *Trends in Biochemical Sciences*, vol. 40, no. 1, pp. 4–7, 2015.

[17] I. Chepelev, G. Wei, Q. Tang, and K. Zhao, "Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq," *Nucleic Acids Research*, vol. 37, no. 16, article e106, 2009.

[18] A. C. Miller, N. D. Obholzer, A. N. Shah, S. G. Megason, and C. B. Moens, "RNA-seq-based mapping and candidate identification of mutations from forward genetic screens," *Genome Research*, vol. 23, no. 4, pp. 679–686, 2013.

[19] R. Piskol, G. Ramaswami, and J. B. Li, "Reliable identification of genomic variants from RNA-seq data," *American Journal of Human Genetics*, vol. 93, no. 4, pp. 641–651, 2013.

[20] E. M. Quinn, P. Cormican, E. M. Kenny et al., "Development of strategies for SNP detection in RNA-Seq data: application to lymphoblastoid cell lines and evaluation using 1000 genomes data," *PLoS ONE*, vol. 8, no. 3, article e58815, 2013.

[21] P. P. Amaral and J. S. Mattick, "Noncoding RNA in development," *Mammalian Genome*, vol. 19, no. 7-8, pp. 454–492, 2008.

[22] A. Bhan, I. Hussain, K. I. Ansari, S. A. M. Bobzean, L. I. Perrotti, and S. S. Mandal, "Bisphenol-A and diethylstilbestrol exposure induces the expression of breast cancer associated long noncoding RNA HOTAIR in vitro and in vivo," *Journal of Steroid Biochemistry and Molecular Biology*, vol. 141, pp. 160–170, 2014.

[23] A. Bhan, I. Hussain, K. I. Ansari, S. Kasiri, A. Bashyal, and S. S. Mandal, "Antisense transcript long noncoding RNA (lncRNA) HOTAIR is transcriptionally induced by estradiol," *Journal of Molecular Biology*, vol. 425, no. 19, pp. 3707–3722, 2013.

[24] M. E. Dinger, P. P. Amaral, T. R. Mercer, and J. S. Mattick, "Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications," *Briefings in Functional Genomics and Proteomics*, vol. 8, no. 6, Article ID elp038, pp. 407–423, 2009.

[25] M. E. Dinger, P. P. Amara, T. R. Mercer et al., "Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation," *Genome Research*, vol. 18, no. 9, pp. 1433–1445, 2008.

[26] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[27] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, "Specific expression of long noncoding RNAs in the mouse brain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 2, pp. 716–721, 2008.

[28] S. Schoeftner and M. A. Blasco, "Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II," *Nature Cell Biology*, vol. 10, no. 2, pp. 228–236, 2008.

[29] Y. Guo, S. Zhao, Q. Sheng et al., "RNAseq by total RNA library identifies additional RNAs compared to poly(A) RNA library," *BioMed Research International*, vol. 2015, Article ID 862130, 9 pages, 2015.

[30] R. Andersson, C. Gebhard, I. Miguel-Escalada et al., "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, pp. 455–461, 2014.

[31] D. Hnisz, B. J. Abraham, T. I. Lee et al., "Super-enhancers in the control of cell identity and disease," *Cell*, vol. 155, no. 4, pp. 934–947, 2013.

[32] J. Lovén, H. A. Hoke, C. Y. Lin et al., "Selective inhibition of tumor oncogenes by disruption of super-enhancers," *Cell*, vol. 153, no. 2, pp. 320–334, 2013.

[33] W. A. Whyte, D. A. Orlando, D. Hnisz et al., "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, vol. 153, no. 2, pp. 307–319, 2013.

[34] Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels, "Three-stage quality control strategies for DNA re-sequencing data," *Briefings in Bioinformatics*, vol. 15, no. 6, Article ID bbt069, pp. 879–889, 2013.

[35] Y. Guo, S. Zhao, Q. Sheng et al., "Multi-perspective quality control of Illumina exome sequencing data using QC3," *Genomics*, vol. 103, no. 5-6, pp. 323–328, 2014.

[36] Y. Guo, S. Zhao, F. Ye, Q. Sheng, and Y. Shyr, "MultiRankSeq: multiperspective approach for RNAseq differential expression analysis and quality control," *BioMed Research International*, vol. 2014, Article ID 248090, 8 pages, 2014.

[37] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, article R36, 2013.

[38] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[39] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced heat map and clustering analysis using heatmap3," *BioMed Research International*, vol. 2014, Article ID 986048, 6 pages, 2014.

[40] B. D. Lehmann, J. A. Bauer, X. Chen et al., "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *The Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, 2011.

[41] X. Chen, J. Li, W. H. Gray et al., "TNBCtype: a subtyping tool for triple-negative breast cancer," *Cancer Informatics*, vol. 11, pp. 147–156, 2012.

[42] M. L. Baglia, Q. Cai, Y. Zheng et al., "Dual specificity phosphatase 4 gene expression in association with triple-negative breast cancer outcome," *Breast Cancer Research and Treatment*, vol. 148, no. 1, pp. 211–220, 2014.

[43] D. C. Koboldt, Q. Zhang, D. E. Larson et al., "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research*, vol. 22, no. 3, pp. 568–576, 2012.

[44] R. M. Durbin, D. L. Altshuler, G. R. Abecasis et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[45] Y. Guo, J. Li, C.-I. Li, J. Long, D. C. Samuels, and Y. Shyr, "The effect of strand bias in Illumina short-read sequencing data," *BMC Genomics*, vol. 13, article 666, 2012.

[46] Y. Guo, J. Long, J. He et al., "Exome sequencing generates high quality data in non-target regions," *BMC Genomics*, vol. 13, no. 1, article 194, 2012.

[47] J. Wang, L. Raskin, D. C. Samuels, Y. Shyr, and Y. Guo, "Genome measures used for quality control are dependent on gene function and ancestry," *Bioinformatics*, vol. 31, no. 3, pp. 318–323, 2015.

[48] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, article e164, 2010.

[49] P. P. Reis, L. Waldron, R. S. Goswami et al., "MRNA transcript quantification in archival samples using multiplexed, color-coded probes," *BMC Biotechnology*, vol. 11, article 46, 2011.

[50] R. A. Roberts, C. M. Sabalos, M. L. LeBlanc et al., "Quantitative nuclease protection assay in paraffin-embedded tissue replicates prognostic microarray gene expression in diffuse large-B-cell lymphoma," *Laboratory Investigation*, vol. 87, no. 10, pp. 979–997, 2007.

[51] T. Nielsen, B. Wallden, C. Schaper et al., "Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens," *BMC Cancer*, vol. 14, article 177, 2014.

[52] A. Sapino, P. Roepman, S. C. Linn et al., "MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue," *Journal of Molecular Diagnostics*, vol. 16, no. 2, pp. 190–197, 2014.

[53] M. D. Burstein, A. Tsimelzon, G. M. Poage et al., "Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer," *Clinical Cancer Research*, vol. 21, no. 7, pp. 1688–1698, 2015.

[54] P. Jézéquel, D. Loussouarn, C. Guérin-Charbonnel et al., "Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response," *Breast Cancer Research*, vol. 17, article 43, 2015.

[55] X. Adiconis, D. Borges-Rivera, R. Satija et al., "Comparative analysis of RNA sequencing methods for degraded or low-input samples," *Nature Methods*, vol. 10, no. 7, pp. 623–629, 2013.

[56] Y. Guo, Q. H. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.

[57] Q. Sheng, S. Zhao, C. Li, Y. Shyr, and Y. Guo, "Practicability of detecting somatic point mutation from RNA high throughput sequencing data," *Genomics*, vol. 107, no. 5, pp. 163–169, 2016.

[58] J. Hedegaard, K. Thorsen, M. K. Lund et al., "Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue," *PLoS ONE*, vol. 9, no. 5, Article ID e98187, 2014.

[59] P. Li, A. Conley, H. Zhang, and H. L. Kim, "Whole-transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq," *BMC Genomics*, vol. 15, article 1087, 2014.

[60] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, "Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling," *BMC Genomics*, vol. 15, article 419, 2014.